Г.Б. Поднебесова

ТЕОРИЯ ИНФОРМАЦИИ. ДАННЫЕ. ЗНАНИЯ

Учебно-практическое пособие

Министерство просвещения РФ
Федеральное государственное бюджетное образовательное
учреждение высшего образования
«Южно-Уральский государственный
гуманитарно-педагогический университет»

Г.Б. Поднебесова

ТЕОРИЯ ИНФОРМАЦИИ. ДАННЫЕ. ЗНАНИЯ

Учебно-практическое пособие

Челябинск
2022

УДК 001.8(021)
ББК 73я73
П 44

Поднебесова, Г.Б. Теория информации. Данные. Знания: учебно-практическое пособие / Г.Б. Поднебесова. – Челябинск: Изд-во Южно-Урал. гос. гуман.-пед. ун-та, 2022. – 112 с. – ISBN 978-5-907611-46-7. – Текст: непосредственный.

Учебно-практическое пособие содержит материал для изучения курса «Теория информации. Данные. Знания». Пособие предназначено для организации аудиторной и самостоятельной работы студентов, обучающихся по направлению «Информационные системы и технологии».

Пособие адресовано преподавателям и учителям, которым интересна данная предметная область.

Рецензенты: С.А. Загребина, д-р физ.-мат. наук, профессор ЮУрГУ
А.А. Рузаков, канд. пед. наук, доцент ЮУрГГПУ

ISBN 978-5-907611-46-7

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Стратегия развития современного общества на основе знаний и высокоэффективных технологий потребовала внесения значительных корректив в педагогическую теорию и практику, активизировала поиск новых моделей образования, направленных на повышение уровня квалификации и профессионализма педагогов, на удовлетворение потребностей общества в специалистах, способных к успешной адаптации.

Учебный курс вводит студентов в современные проблемы теоретической информатики. Основной акцент в курсе делается на методологические аспекты и математический аппарат информатики, составляющие ядро широкого спектра научно-технических и социально-экономических информационных технологий, которые реально используются современным мировым профессиональным сообществом в теоретических исследованиях и практической деятельности.

Учебная дисциплина «Теория информации, данные, знания», основывается на материале предшествующих ей дисциплин Математики (Математический анализ, Алгебра и теория чисел), курса Абстрактной и компьютерной алгебры.

Программа курса предусматривает аудиторные занятия (лекции и практические занятия – лабораторные практикумы) и самостоятельную работу студентов. В самостоятельную работу студентов входит освоение теоретического материала, выполнение индивидуальных заданий, подготовка сообщений и написание реферата по разделам дисциплины.

В результате изучения дисциплины студент должен:

1) иметь представление об общих проблемах и задачах теоретической информатики;

2) иметь представление об основных принципах и этапах информационных процессов;

3) знать наиболее широко используемые классы информационных моделей и основные математические методы получения, хранения, обработки, передачи и использования информации;

4) уметь применять математический аппарат анализа и синтеза информационных систем;

5) уметь применять методы программирования и навыки работы с математическими пакетами для решения практических задач хранения и обработки информации.

В данном случае оптимальной формой контроля усвоения материала становится рейтинговая система оценки учебных достижений студентов, а именно:

– текущий контроль: практические занятия и тесты;

– промежуточный контроль: индивидуальные задания;

– итоговый контроль: экзамен в виде теста.

Дисциплина относится к модулю обязательной части Блока 1 «Дисциплины/модули» основной профессиональной образовательной программы по направлению подготовки 09.03.02 «Информационные системы и технологии» (уровень образования бакалавр).

Таблица 1 – Перечень планируемых результатов обучения

№ п/п	Код и наименование компетенции по ФГОС
Код и наименование индикатора достижения компетенции	
1	ОПК-1. Способен применять естественнонаучные и общеинженерные знания, методы математического анализа и моделирования, теоретического и экспериментального исследования в профессиональной деятельности
	ОПК.1.1. Знать основы математики, физики, вычислительной техники и программирования
	ОПК.1.2.Уметь решать стандартные профессиональные задачи с применением естественнонаучных и общеинженерных знаний, методов математического анализа и моделирования
	ОПК.1.3. Иметь навыки теоретического и экспериментального исследования объектов профессиональной деятельности
2	УК-6. Способен управлять своим временем, выстраивать и реализовывать траекторию саморазвития на основе принципов образования в течение всей жизни
	УК.6.1. Знать основные приемы эффективного управления собственным временем; основные методики самоконтроля, саморазвития на протяжении всей жизни
	УК.6.2. Уметь эффективно планировать и контролировать собственное время; использовать методы саморегуляции, саморазвития и самообучения
	УК.6.3. Владеть методами управления собственным временем; технологиями приобретения, использования и обновления социокультурных и профессиональных знаний, умений и навыков; методиками саморазвития и самообразования в течение всей жизни

Таблица 2 – Образовательные результаты по дисциплине

№ п/п	Код и наименование индикатора достижения компетенции	Образовательные результаты по дисциплине
1	ОПК.1.1. Знать основы математики, физики, вычислительной техники и программирования	З.1. Иметь представление об общих проблемах и задачах теории информации З.2. Способы кодирования информации
2	ОПК.1.2. Уметь решать стандартные профессиональные задачи с применением естественнонаучных и общеинженерных знаний, методов математического анализа и моделирования	У.1. Определять количество информации У.2. Применять алгоритмы для кодирования информации
3	ОПК.1.3. Иметь навыки теоретического и экспериментального исследования объектов профессиональной деятельности	В.1. Владеть навыками кодирования информации
1	УК.6.1. Знать основные приемы эффективного управления собственным временем; основные методики самоконтроля, саморазвития на протяжении всей жизни	З.3. Наиболее широко используемые классы информационных моделей и основные математические методы получения, хранения, обработки, передачи и использования информации
2	УК.6.2. Уметь эффективно планировать и контролировать собственное время; использовать методы саморегуляции, саморазвития и самообучения	У.3. Применять методы получения, хранения, обработки информации при разработке ИС

Общепрофессиональной компетенцией, формируемой в результате освоения дисциплины, является способность применять естественнонаучные и общеинженерные знания, методы математического анализа и моделирования, теоретического и экспериментального исследования в профессиональной деятельности (ОПК-1). Универсальная компетенция, формируемая в процессе изучения данного курса, – способность управлять своим временем, выстраивать и реализовывать траекторию саморазвития на основе принципов образования в течение всей жизни (УК-6).

СОДЕРЖАНИЕ РАЗДЕЛОВ ДИСЦИПЛИНЫ

1. ИНФОРМАЦИЯ

Роль информации в современном обществе. Виды информационных процессов. Принципы получения, хранения, обработки и использования информации. Данные, знания. Свойства декларативных знаний. Классификация знаний. Модели представления знаний. Канал связи. Характеристики каналов связи. Условия оптимального использования каналов связи. Моделирование данных. Сетевая теория информации.

2. КОЛИЧЕСТВО ИНФОРМАЦИИ

Подходы к измерению количества информации. Метод Хартли. Статистический подход. Семантический подход. Прагматический подход. Алфавиты. Системы счисления. Кодирование. Количество информации. Избыточность. Виды избыточности. Передача дискретных сообщений по каналу без шумов и с шумами.

3. ТЕОРИЯ КОДИРОВАНИЯ

Алфавитное кодирование. Разделимые коды. Префиксные коды. Критерий однозначности декодирования. Неравенство МакМиллана для разделимых кодов. Оптимальные коды. Методы построения оптимальных кодов. Метод Хаффмана. Самокорректирующиеся коды. Основные принципы сжатия информации. Сжатие с потерями и без потерь. Арифметический и вероятностный методы. Криптография. Электронная подпись. Методы защиты данных. Хэш-функции. Электронная цифровая подпись. Криптосистема Эль-Гамаля.

4. ПОМЕХОУСТОЙЧИВОЕ КОДИРОВАНИЕ

Помехи и их источники. Классификация помехоустойчивых кодов. Коды Хэмминга. Коды Хэмминга, исправляющие одиночную ошибку. Помехоустойчивое кодирование в системах сотовой связи. Виды помех и борьба с ними в системах сотовой связи. Стандарты сотовой связи. Мировые информационные ресурсы и глобальные информационные сети. Классификация мировых информационных ресурсов.

ТЕМЫ И ПЛАНЫ ЛАБОРАТОРНЫХ ЗАНЯТИЙ

Модуль 1. Информация

1. Вычисление статистических характеристик текстовой информации (2 часа):

1) определение количества информации;

2) построение таблицы частот;

3) анализ частоты появления букв русского алфавита в тексте с помощью Excel.

2. Работа с псевдослучайными числами и оценка их качества статистическими тестами (2 часа):

1) применение функций Excel для получения случайных чисел;

2) работа с программой Псевдослучайные последовательности;

3) рассмотрение систем оценки качества генераторов псевдослучайных последовательностей.

3. Разработка программы формирования псевдослучайных чисел и оценка их качества статистическими тестами (4 часа):

1) разработка программы на C# для оценки качества полученных последовательностей;

2) проверка на равномерность распределения.

4. Моделирование данных (4 часа):

1) Подготовка сообщение по одной из тем.

2) Разработка и защита интеллект карты.

Модуль 2. Количество информации

1. Разработка программы подсчета количества информации методом К. Шеннона (2 часа):

1) метод К. Шеннона;

2) реализация алгоритма вычисления количества информации;

2. Разработка программы подсчета количества информации методом Р. Хартли (2 часа):

1) метод Р. Хартли;

2) реализация алгоритма вычисления количества информации.

3. Разработка программы подсчета количества прагматической информации (2 часа):

1) метод А.А. Харкевича;

2) реализация алгоритма вычисления количества прагматической информации.

Модуль 3. Кодирование

1. Кодирование и декодирование символьной информации с использованием различных кодовых таблиц (2 часа):

1) различные кодовые таблицы;

2) кодовые таблицы ASCII;

3) кодирование символьной информации.

2. Кодирование графической, звуковой и видео информации (2 часа):

1) двоичное кодирование графической информации;

2) дискретизация и квантование;

3) кодирование видеоинформации. Метод JPEG. Алгоритм MPEG;

4) примеры кодирования звуковой и видео информации.

3. Кодирование информации методами Шеннона-Фано и Хаффмана (4 часа):

1) кодирование информации;

2) метод Шеннона–Фано;

3) метод Хаффмана.

4. Сжатие данных по методу Лемпеля–Зива (2 часа):

1) LZ-метод;

2) алгоритм LZ78.

Модуль 4. Помехоустойчивое кодирование

1. Обнаружение одиночной ошибки методом Хемминга (2 часа):

– обнаружение одиночной ошибки (2 метода).

2. Защита реферата (помехоустойчивые коды) (2 часа).

3. Алгоритмы цифровой подписи (2 часа):

1) шифр Эль-Гамаля. Описание метода;

2) алгоритм цифровой подписи DSA. Пример.

МОДУЛЬ 1. ТЕОРИЯ ИНФОРМАЦИИ

Лабораторная работа 1
Вычисление статистических характеристик текстовой информации

Теоретические сведения

Важными характеристиками текста являются повторяемость букв, пар букв (биграмм) и вообще m-ок (m-грамм), сочетаемость букв друг с другом, чередование гласных и согласных, и некоторые другие. Эти характеристики являются достаточно устойчивыми (см. Приложение 4).

Для русского языка частоты знаков алфавита, в котором отождествлены Е с Ё, Ь с Ъ, а также имеется знак пробела между словами, приведены в таблице 2.

Таблица 1.1 – Частоты знаков алфавита

Символ / Вероятность		**Символ / Вероятность**		**Символ / Вероятность**		**Символ / Вероятность**	
пробел	0.175	О	0.090	Е, Ё	0.072	А	0.062
И	0.062	Т	0.053	Н	0.053	С	0.045
Р	0.040	В	0.038	Л	0.035	К	0.028
М	0.026	Д	0.025	П	0.023	У	0.021
Я	0.018	Ы	0.016	З	0.016	Ь, Ъ	0.014
Б	0.014	Г	0.013	Ч	0.012	Й	0.010
Х	0.009	Ж	0.007	Ю	0.006	Ш	0.006
Ц	0.004	Щ	0.003	Э	0.003	Ф	0.002

Некоторая разница значений частот в различных источниках объясняется тем, что частоты существенно зависят не только от длины текста, но и от его характера.

Если бы сообщения передавались с помощью равновероятных букв алфавита и между собой статистически независимых, то энтропия таких сообщений была бы максимальной. На самом деле реальные сообщения строятся из не равновероятных букв алфавита с наличием статистических связей между буквами. Поэтому энтропия реальных сообщений – $H_р$, оказывается много меньше оптимальных сообщений – $H_о$. Допустим, нужно передать сообщение, содержащее количество информации, равное I. Источнику, обладающему энтропией на букву, равной $H_р$, придется затратить некоторое число $n_р$, то есть $I = n_р H_р$.

Если энтропия источника была бы H_0, то пришлось бы затратить меньше букв на передачу этого же количества информации $I = n_0 H_0$, т.е.

$$n_o = \frac{1}{H_o} < n_p . \qquad (1)$$

Таким образом, часть букв $n_р - n_о$ является как бы лишними, избыточными. Мера удлинения реальных сообщений по сравнению с оптимально закодированными и представляет собой избыточность D.

$$D = 1 - \frac{H_p}{H_o} = 1 - \frac{n_o}{n_p} = \frac{n_p - n_o}{n_o} . \qquad (2)$$

Но наличие избыточности нельзя рассматривать как признак несовершенства источника сообщений. Наличие избыточности (2) способствует повышению помехоустойчивости сообщений. Высокая избыточность естественных языков обеспечивает надежное общение между людьми.

Задание 1. Определить количество информации (по Хартли), содержащееся в заданном сообщении, при условии, что значениями являются буквы кириллицы.

«Информация в общем виде является свойством материальных объектов, существует вечно, никогда не возникала и никогда не исчезает.».

Задание 2. Построить таблицу распределения частот символов, характерных для заданого сообщения (см. Приложение 4). Производится так называемая частотная селекция, текст сообщения анализируется как поток символов и высчитывается частота встречаемости каждого символа. Сравнить с имеющимися данными в таблице 2.

Задание 3. На основании полученных данных определить среднее и полное количество информации, содержащееся в заданном сообщении.

Задания для самостоятельной работы

Оценить избыточность сообщения из задания 3.

Лабораторная работа 2
Разработка программы формирования псевдослучайных чисел и оценка их качества статистическими тестами

Теоретические сведения

Генератор случайных чисел (ГСЧ) должен выдавать близкие к следующим значения статистических параметров, характерных для равномерного случайного закона:

$$m_r = \frac{\sum_{i=1}^{n} r_i}{n} \approx 0.5$$ – математическое ожидание

$$D_r = \frac{\sum_{i=1}^{n} (r_i - m_r)^2}{n} \approx 0.0833$$ – дисперсия

$$\sigma_r = \sqrt{D_r} \approx 0.2887$$ – среднеквадратичное отклонение

Рис. 1.1 – Числовые характеристики случайной величины

Частотный тест позволяет выяснить, сколько чисел попало в интервал (m_r – σ_r; m_r + σ_r), то есть (0.5 – 0.2887; 0.5 + 0.2887) или, в конечном итоге, (0.2113; 0.7887). Так как 0.7887 – 0.2113 = 0.5774, заключаем, что в хорошем ГСЧ в этот интервал должно попадать около 57.7% из всех выпавших случайных чисел (см. рис. 2).

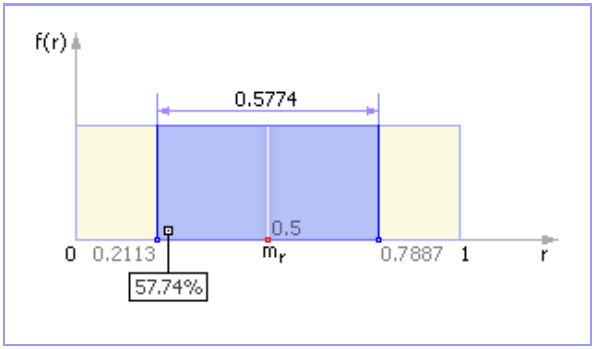


Рис. 1.2 – Частотная диаграмма идеального ГСЧ в случае проверки его на частотный тест

Также необходимо учитывать, что количество чисел, попавших в интервал (0; 0,5), должно быть примерно равно количеству чисел, попавших в интервал (0,5; 1).

Задание 1. Получить случайные последовательности чисел в Excel.

Использовать функции СЛЧИС и СЛУЧМЕЖДУ (например, СЛЧИС (100 - 1) + 1), СЛУЧМЕЖДУ (1; 100)).

Таблица 1.2 – Пример оформления таблицы

№	случ1	случ2	случ3	(m-r)^2	случ4	(m-r)^2
1	0,497	1	15,707	0,014	0,231	0,086
2	0,005	0	5,169	0,139	0,432	0,009
3	0,085	0	10,097	0,085	0,543	0
4	0,525	0	3,89	0,022	0,865	0,116
5	0,625	0	19,255	0,061	0,123	0,161
6	0,067	1	3,937	0,096	0,142	0,146
7	0,178	1	1,106	0,04	0,655	0,017
8	0,07	1	6,722	0,095	0,543	0
9	0,402	1	16,681	0,001	0,695	0,029
10	0,28	0	4,518	0,01	0,958	0,188
11	0,528	1	17,048	0,023	0,764	0,058
12	0,282	1	19,336	0,009	0,362	0,026
13	0,8	1	13,2	0,179	0,764	0,058
14	0,14	0	9,414	0,056	0,154	0,137
15	0,842	0	19,154	0,216	0,874	0,122
16	0,586	1	3,542	0,043	0,433	0,008
17	0,482	1	16,36	0,011	0,436	0,008
18	0,207	1	14,188	0,029	0,976	0,204
19	0,833	1	19,727	0,207	0,217	0,094
20	0,118	0	18,788	0,068	0,318	0,043
Сумма	7,552	12	237,838	1,403	10,487	1,511
m				0,378		0,524
Дисперсия				0,07		0,076
Сигма				0,265		0,275
Разница						0,01

Задание 2. Откройте программу «Псевдослучайные последовательности», запустив файл **RNS.exe.**

1. Выберите один из методов построения псевдослучайных последовательностей.

2. Изучите интерфейс программы.

a. Какие данные предлагаются ввести?

b. За что отвечает кнопка Контроль данных**?**

c. Какие вкладки можно просмотреть для каждой последовательности?

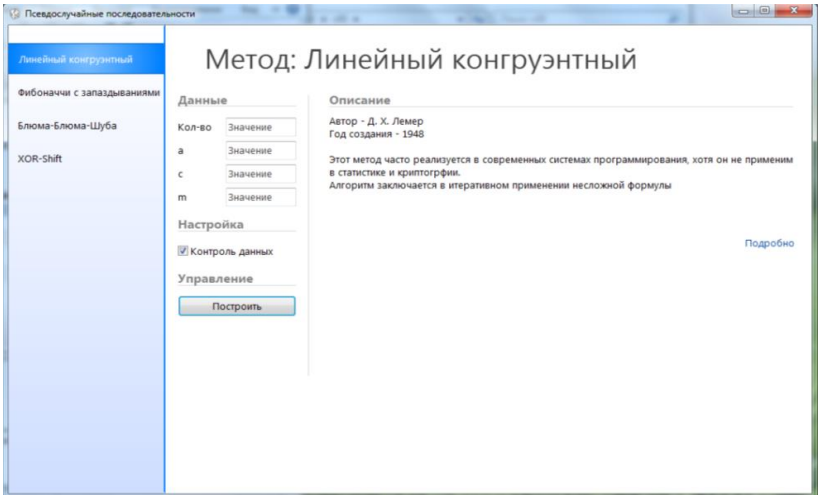


Рис. 1.3 – Окно метода Линейны конгруэнтный

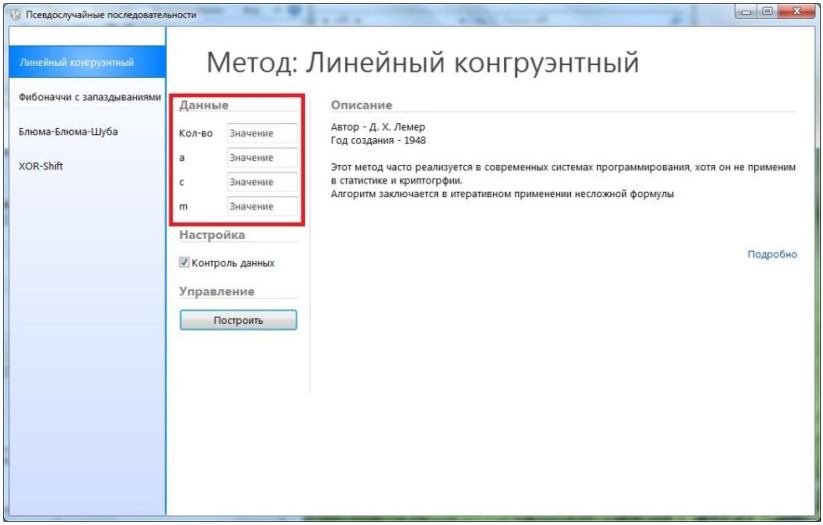


Рис. 1.4 – Данные метода

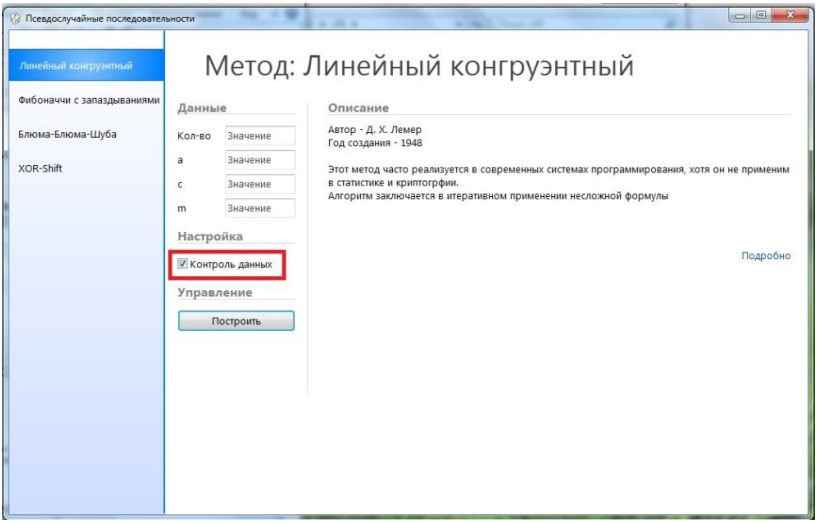


Рис. 1.5 – Контроль данных

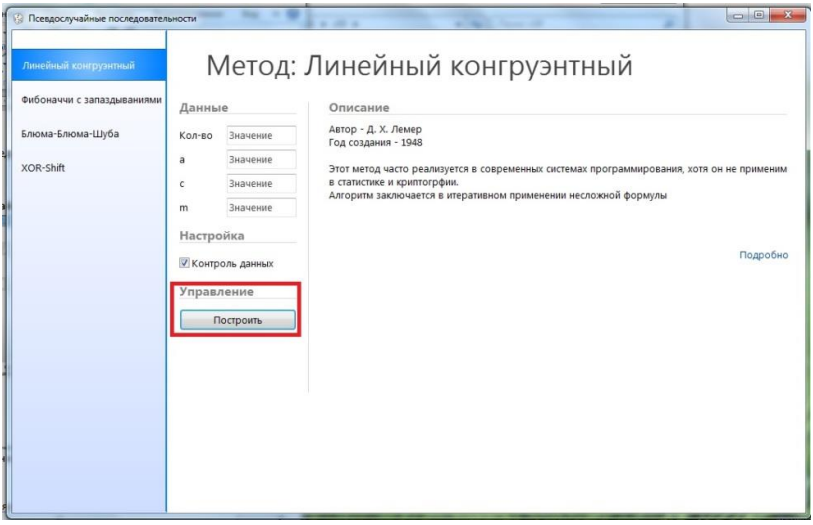


Рис. 1.6 – Управление построением диаграмм

3. Изучите справку программы (перейти по ссылке Подробно).

a. Ознакомьтесь с общей информацией о программе.

b. Какая информация содержится на вкладке Методы?

4. Постройте последовательность из 50 элементов, пользуясь Линейным конгруэнтным методом.

a. Что можно сказать о повторениях в ряду последовательности?

b. Перейдите на вкладку Статистика, в каком интервале больше всего элементов?

c. Что происходит, когда вводятся некорректные данные?

5. Постройте последовательности по всем методам.

Задание 3

1. Используя программу «Псевдослучайные последовательности» определить, чему равна длина периода последовательности, полученной линейным конгруэнтным методом с параметрами a = 69069, c = 19, m = 256?

2. Проверить соблюдение правила для равномерного распределения псевдослучайных чисел: среднее значение элементов = $^1/_2$, дисперсия = $^1/_{12}$ для всех методов программы.

3. Используя программу «Псевдослучайные последовательности», вычислить дисперсию и среднее отклонение для построенной последовательности:

a. Методом XOR-Shift, N = 20 (N – количество элементов).

b. Методом Фибоначчи с запаздываниями, N = 25.

4. Провести графический тест – Проверка на монотонность.

a. За исходную последовательность взять последовательность, построенную Линейным конгруэнтным методом, n = 100.

b. За исходную последовательность взять последовательность, построенную методом Фибоначчи с запаздываниями, n = 100.

5. Провести следующие оценочные тесты:

Частотный тест.

Пользоваться вкладкой Отчет программы «Псевдослучайные последовательности». За исходную последовательность брать последовательности, построенные различными генераторами, реализованными в программе «Псевдослучайные последовательности»

Лабораторные работы 3–4
Разработка программы формирования псевдослучайных чисел

Теоретические сведения

Частотный тест позволяет выяснить, сколько чисел попало в интервал $(m_r - \sigma_r; m_r + \sigma_r)$, то есть $(0.5 - 0.2887; 0.5 + 0.2887)$ или, в конечном итоге, $(0.2113; 0.7887)$. Так как $0.7887 - 0.2113 = 0.5774$, заключаем, что в хорошем ГСЧ в этот интервал должно попадать около 57.7% из всех выпавших случайных чисел (см. рис. 1).

Также необходимо учитывать, что количество чисел, попавших в интервал $(0; 0.5)$, должно быть примерно равно количеству чисел, попавших в интервал $(0.5; 1)$.

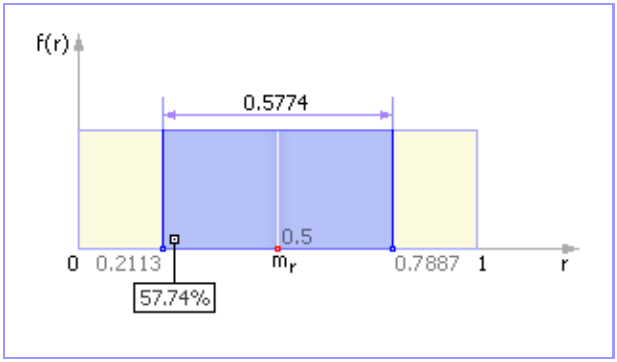


Рис. 1.7 – Частотная диаграмма идеального ГСЧ
в случае проверки его на частотный тест

Критерий «хи-квадрат» (χ^2-критерий) – это один из самых известных статистических критериев; он является основным методом, используемым в сочетании с другими критериями. Критерий «хи-квадрат» был предложен в 1900 году Карлом Пирсоном. Его замечательная работа рассматривается как фундамент современной математической статистики.

Для нашего случая проверка по критерию «хи-квадрат» позволит узнать, насколько созданный нами *реальный* ГСЧ близок к эталону ГСЧ, то есть удовлетворяет ли он требованию равномерного распределения или нет.

Частотная диаграмма *эталонного* ГСЧ представлена на рис. 1.8. Так как закон распределения эталонного ГСЧ равномерный, то (теоретическая) вероятность p_i попадания чисел в i-ый интервал (всего этих интервалов k) равна $p_i = 1/k$. И, таким образом, в каждый из k интервалов попадет *ровно* по $p_i \cdot N$ чисел (N — общее количество сгенерированных чисел).

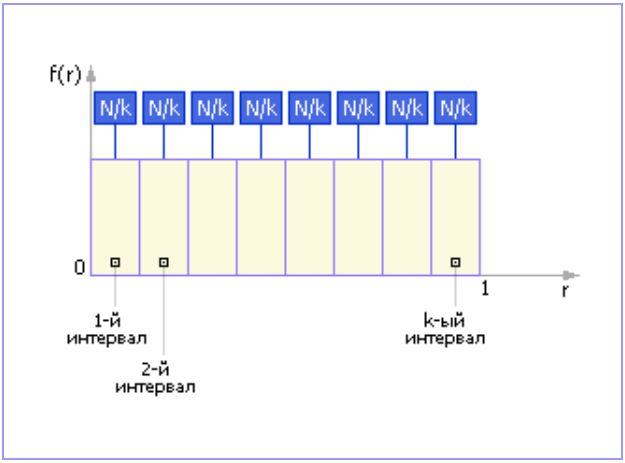


Рис. 1.8 – Частотная диаграмма эталонного ГСЧ

Реальный ГСЧ будет выдавать числа, распределенные (причем, не обязательно равномерно!) по k интервалам и в каждый интервал попадет по n_i чисел (в сумме $n_1 + n_2 + \ldots + n_k = N$). Как же нам определить, насколько испытываемый ГСЧ хорош и близок к эталонному? Вполне логично рассмотреть квадраты разностей между полученным количеством чисел n_i и «эталонным» $p_i \cdot N$. Сложим их, и в результате получим:

$$\chi^2_{\text{эксп.}} = (n_1 - p_1 \cdot N)^2 + (n_2 - p_2 \cdot N)^2 + \ldots + (n_k - p_k \cdot N)^2.$$

Из этой формулы следует, что чем меньше разность в каждом из слагаемых (а значит, и чем меньше значение $\chi^2_{\text{эксп.}}$), тем сильнее закон распределения случайных чисел, генерируемых реальным ГСЧ, тяготеет к равномерному.

В предыдущем выражении каждому из слагаемых приписывается одинаковый вес (равный 1), что на самом деле может не соответствовать действительности; поэтому

для статистики «хи-квадрат» необходимо провести нормировку каждого i-го слагаемого, поделив его на $p_i \cdot N$:

$$\chi^2_{\text{эксп.}} = \frac{(n_1 - p_1 \cdot N)^2}{p_1 \cdot N} + \frac{(n_2 - p_2 \cdot N)^2}{p_2 \cdot N} + \ldots + \frac{(n_k - p_k \cdot N)^2}{p_k \cdot N}$$

Наконец, запишем полученное выражение более компактно и упростим его:

$$\chi^2_{\text{эксп.}} = \sum_{i=1}^{k} \frac{(n_i - p_i \cdot N)^2}{p_i \cdot N} = \frac{1}{N} \sum_{i=1}^{k} \left(\frac{n_i^2}{p_i} \right) - N$$

Получили значение критерия «хи-квадрат» для *экспериментальных* данных.

В табл. 1 приведены *теоретические* значения «хи-квадрат» ($\chi^2_{\text{теор.}}$), где $\nu = N - 1$ – это число степеней свободы, **p** – это доверительная вероятность, задаваемая пользователем, который указывает, насколько ГСЧ должен удовлетворять требованиям равномерного распределения, или **p** – это вероятность того, что экспериментальное значение $\chi^2_{\text{эксп.}}$ будет меньше табулированного (теоретического) $\chi^2_{\text{теор.}}$ или равно ему.

Приемлемым считают **p** от 10% до 90%.

Если $\chi^2_{\text{эксп.}}$ лежит в некотором диапазоне, между двумя значениями $\chi^2_{\text{теор.}}$, которые соответствуют, например, **p** = 25% и **p** = 50%, то можно считать, что значения случайных чисел, порождаемые датчиком, вполне являются случайными.

При этом дополнительно надо иметь в виду, что все значения $p_i \cdot N$ должны быть достаточно большими, например, больше 5 (выяснено эмпирическим путем). Только тогда (при достаточно большой статистической выборке) условия проведения эксперимента можно считать удовлетворительными.

Таблица 1.3 – Некоторые процентные точки χ^2-распределения

	p = 1%	**p = 5%**	**p = 25%**	**p = 50%**	**p = 75%**	**p = 95%**	**p = 99%**
ν = 1	0.00016	0.00393	0.1015	0.4549	1.323	3.841	6.635
ν = 2	0.02010	0.1026	0.5754	1.386	2.773	5.991	9.210
ν = 3	0.1148	0.3518	1.213	2.366	4.108	7.815	11.34
ν = 4	0.2971	0.7107	1.923	3.357	5.385	9.488	13.28
ν = 5	0.5543	1.1455	2.675	4.351	6.626	11.07	15.09
ν = 6	0.8721	1.635	3.455	5.348	7.841	12.59	16.81
ν = 7	1.239	2.167	4.255	6.346	9.037	14.07	18.48
ν = 8	1.646	2.733	5.071	7.344	10.22	15.51	20.09
ν = 9	2.088	3.325	5.899	8.343	11.39	16.92	21.67
ν = 10	2.558	3.940	6.737	9.342	12.55	18.31	23.21
ν = 11	3.053	4.575	7.584	10.34	13.70	19.68	24.72
ν = 12	3.571	5.226	8.438	11.34	14.85	21.03	26.22
ν = 15	5.229	7.261	11.04	14.34	18.25	25.00	30.58
ν = 20	8.260	10.85	15.45	19.34	23.83	31.41	37.57
ν = 30	14.95	18.49	24.48	29.34	34.80	43.77	50.89
ν = 50	29.71	34.76	42.94	49.33	56.33	67.50	76.15
ν > 30	$\nu + \mathrm{sqrt}(2\nu) \cdot x_p + 2/3 \cdot x^2_p - 2/3 + O(1/\mathrm{sqrt}(\nu))$						
x_p =	–2.33	–1.64	–0.674	0.00	0.674	1.64	2.33

Итак, процедура проверки имеет следующий вид.

1. Диапазон от 0 до 1 разбивается на k равных интервалов.

2. Запускается ГСЧ N раз (N должно быть велико, например, $N/k > 5$).

3. Определяется количество случайных чисел, попавших в каждый интервал: n_i, $i = 1, \ldots, k$.

4. Вычисляется экспериментальное значение $\chi^2_{эксп.}$ по следующей формуле:

$$\chi^2_{эксп.} = \sum_{i=1}^{k} \frac{(n_i - p_i \cdot N)^2}{p_i \cdot N} = \frac{1}{N} \sum_{i=1}^{k} \left(\frac{n_i^2}{p_i} \right) - N,$$

где $p_i = 1/k$ — теоретическая вероятность попадания чисел в k-ый интервал.

5. Путем сравнения экспериментально полученного значения $\chi^2_{эксп.}$ с теоретическим $\chi^2_{теор.}$ (из табл. 1.3) делается вывод о пригодности генератора для использования. Для этого:

а) входим в табл. 1.3 (**строка = количество экспериментов – 1**);

б) сравниваем вычисленное $\chi^2_{эксп.}$ с $\chi^2_{теор.}$, встречающимися в строке.

При этом возможно три случая.

Первый случай: $\chi^2_{эксп.}$ много больше любого $\chi^2_{теор.}$ в строке – гипотеза о случайности равномерного генератора не выполняется (разброс чисел слишком велик, чтобы быть случайным).

Второй случай: $\chi^2_{эксп.}$ много меньше любого $\chi^2_{теор.}$ в строке – гипотеза о случайности равномерного генератора не выполняется (разброс чисел слишком мал, чтобы быть случайным).

Третий случай: $\chi^2_{эксп.}$ лежит между значениями $\chi^2_{теор.}$ двух рядом стоящих столбцов – гипотеза о случайности равномерного генератора выполняется с вероятностью **p** (то есть в **p** случаях из 100).

Заметим, что чем ближе получается **p** к значению 50%, тем лучше.

Задание 1. Напишите программу на C# для оценки качества полученных последовательностей.

1) ГСЧ должен выдавать близкие к следующим значения статистических параметров, характерных для равномерного случайного закона:

$$m_r = \frac{\sum_{i=1}^{n} r_i}{n} \approx 0.5$$ – математическое ожидание;

$$D_r = \frac{\sum_{i=1}^{n} (r_i - m_r)^2}{n} \approx 0.0833$$ – дисперсия;

$$\sigma_r = \sqrt{D_r} \approx 0.2887$$ – среднеквадратичное отклонение.

Лабораторные работы 5–6

Моделирование данных

Теоретические сведения

Этапы информационного моделирования:

1. Выбор объекта моделирования;
2. Определение цели моделирования;
3. Системный анализ объекта моделирования;
4. Построение информационной модели;
5. Создание компьютерной модели;
6. Использование компьютерной модели.

Задание 1

1. Разработать интеллект карту по теме «Моделирование данных». Воспользуемся редактором https://www.mindmeister.com/ (рис. 1). Для разработки можно использовать любой редактор ментальных карт.

2. Выбрать одну из предложенных тем. Например, Модели дискретного канала. Описать каждый из каналов – Симметричный канал без памяти, Марковский канал и др. (не менее 4).

3. Провести системный анализ построенной модели (рис.1.10).

4-5. Создать информационную и компьютерную модель дискретного канала.

6. Защитить созданную интеллект-карту.

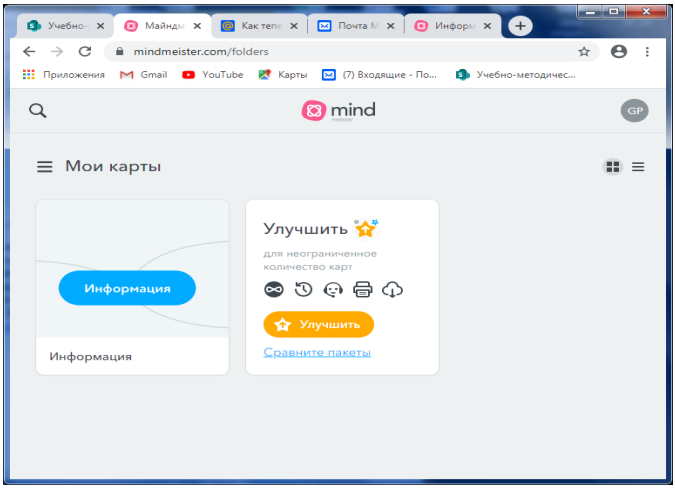


Рис. 1.9 – Главное окно конструктора Интеллект-карт Mindmeister

Индивидуальная работа 1

Анализ текстов

Выполнить анализ текста на иностранном языке объемом не менее 500 слов (см. Приложение 5).

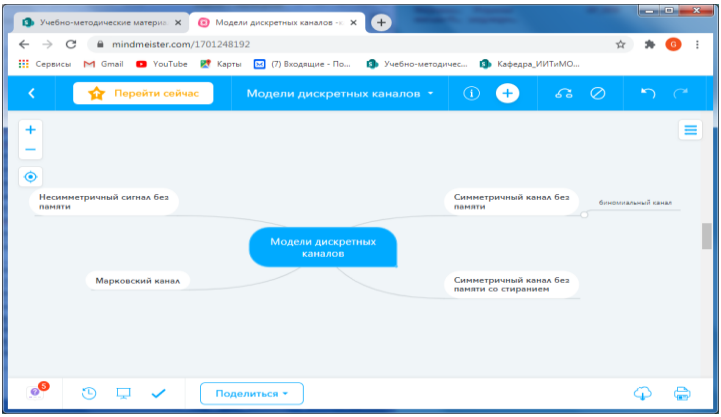


Рис. 1.10 – Пример интеллект карты

Индивидуальная работа 2
Моделирование данных

Разработать интеллект карту по одной из предложенных тем.

Таблица 1.4 – Темы для разработки интеллект-карт

1	Кодирование сети	11	Rate Distortion Theory
2	Случайное сетевое кодирование	12	Колмогорова сложность
3	Канал с множественным доступом	13	Теория информации в физике
4	Широковещательный канал	14	Теория информации, связанная с геномом
5	Relay Channel	15	Кодирование канала
6	Interference channel	16	Синергетическая теория информации
7	Когнитивное радио	17	Квантовая теория информации
8	Масштабирование сети	18	Модели дискретных каналов
9	Портфельная теория	19	Адаптивные системы передачи данных
10	Универсальное кодирование источника	20	Спутниковые каналы связи

Вопросы к модулю 1

1. Дать определение понятия «информация».
2. В чем отличие между данными и знаниями?
3. Категории свойств информации.
4. В каких формах представляется информация?
5. В чем заключается смысл теоремы В.А. Котельникова (Г. Найквиста)?
6. Как вычислить объем сигнала?
7. Чему равна ёмкость канала связи?
8. Что такое энтропия?
9. Чему равна условная энтропия?
10. Как вычислить энтропию сложного опыта?
11. Дать определение кодирования, кода.
12. Дать определение алфавита, системы счисления.
13. Какая система счисления должна использоваться в компьютерах с точки зрения минимизации?
14. Дать определение информационной избыточности.
15. Какие существуют виды избыточности?

МОДУЛЬ 2. КОЛИЧЕСТВО ИНФОРМАЦИИ

Лабораторная работа 7–9

Разработка алгоритмов подсчета количества информации разными методами

Теоретические сведения

Теоретический материал находится на портале университета (https://cspu.sharepoint.com/Education/Shared%20Documents/Forms/AllItems.aspx?id=%2FEducation%2FShared%20Documents%2FIITiMOI%2FISiT%2F2k%202020%2FTI%5FD%5FKn&viewid=c4d1b0a4%2D2b55%2D4e5b%2D9042%2D6d7bb4925514).

Задание 1. Подготовить форму для подсчета количества информации разными методами (см. рис. 1).

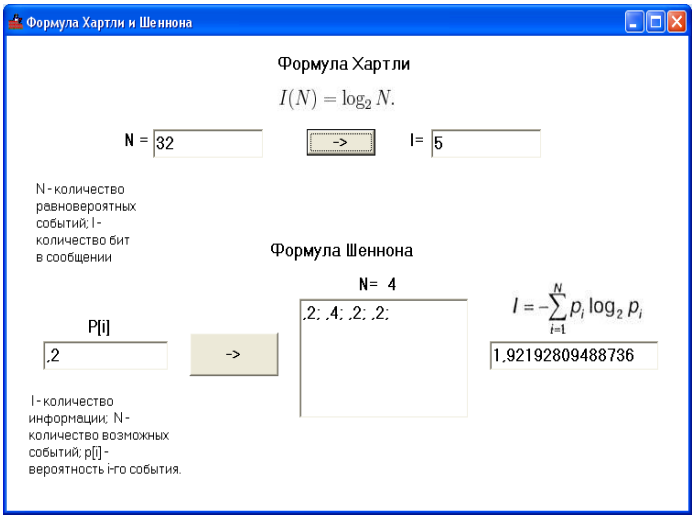


Рис. 2.1 – Формула Хартли и Шеннона

Процедуры для вычисления количества информации приведены ниже.

```
procedure TForm1.Button1Click(Sender: TObject);
  begin
    if ((strtofloat(edit1.Text)>1) or
(strtofloat(edit1.Text)<=0)) then ShowMessage('Вероятность не
может быть больше единицы или меньше либо равна нуль')
                              else
                                begin
                                  x:=ln(strtofloat(edit1.Text))/ln(2);
                                  label3.Caption:= inttostr(strtoint (la-
                                  bel3.Caption)+1);
                                  memo1.Lines.Text:=memo1.Lines.T
                                  ext+ edit1.Text+'; ';
                                  edit2.Text:=floattostr(strtofloat(edit
                                  2.Text)+(strtofloat(edit1.Text)*(-x))) ;
                                end;
end;

procedure TForm1.Button2Click(Sender: TObject);
  begin
    if strtofloat(edit3.text)<1 then ShowMessage('Количество не мо-
    жет быть меньше одного, что это тогда за количество? :)')
                          else  edit4.text:=floattostr(ln(strtofloat(edit3.
                          Text))/ln(2));
end;
```

Задание 2. Рассмотреть метод вычисления количества информации, предложенный А.А. Харкевичем. Реализовать алгоритм.

Задания для самостоятельной работы

Реализовать еще один метод вычисления количества информации.

Вопросы к модулю 2

1. Как связаны понятия неопределенности и вероятности?
2. Как влияет выбор вариантов решения на неопределенность?
3. Чему равно количество информации по Р. Хартли?
4. Как связана вероятность с информативностью?
5. Дать определение количества информации.
6. Как вычисляется количество информации по К. Шеннона?
7. В каком случае формула Шеннона переходит в формулу Хартли?
8. К какому подходу относятся методы Р. Хартли и К. Шеннона?
9. Суть концепции разнообразия Эшби.
10. Как определяется количество информации по А.Н. Колмогорову?
11. Кто предложил связывать понятие «информации» с разнообразием?
12. Зависит ли информация от нашего сознания (по В.М. Глушкову)?
13. Чем занимается наука семиотика?
14. С каких позиций рассматриваются знаковые системы?
15. В чем состоит основная идея семантической концепции информации?
16. В чем заключается подход Р. Карнапа и И. Бар-Хиллела?
17. Кому принадлежит идея учета «запаса знаний»?
18. В чем суть прагматического подхода к измерению количества информации?

МОДУЛЬ 3. КОДИРОВАНИЕ

Лабораторная работа 10

Кодирование и декодирование символьной Информации с использованием различных кодовых таблиц

Теоретические сведения

Таблица символов ASCII

(https://snipp.ru/handbk/table-ascii)

Windows-1251 являлась изменением таблицы ASCII, в которую добавили буквы кириллицы

Unicode

Международная таблица кодировки Unicode, включающая в себя как символы английского, русского, немецкого, арабского и других языков. На каждый символ в такой таблице отводится 16 бит, то есть она позволяет кодировать 65536 символов. Однако использование такой таблицы сильно «утяжеляет» текст.

Задание 1

Закодировать фразу «Теория информации, данные, знания».

MS-DOS, КОИ-8, ISO, Mac и другие

https://www.i5t.ru/images/pdf_files/informatika/kodirovki.pdf

Задание 2

Раскодировать с помощью кодовой таблицы КОИ-7:

165 160 160 160 168

161 237 230 233 161

171 160 160 160 174

Задание 3. Перевести текст:

The Unicode Standard is a universally recognized coding system for more than 120,000 characters, using either 8-bit (UTF-8) or 16-bit (UTF-16) encoding. This chart shows the character symbol and the corresponding hexadecimal UTF-8 code. For the first 127 characters, UTF-8 and ASCII are identical.

EBSCO illustration.

Unicode defines 1,114,112 code points. Each code point is assigned a hexadecimal number ranging from 0 to 10FFFF. When written, these values are typically preceded by U+. For example, the letter *J* is assigned the hexadecimal number 004A and is written U+004A. The Unicode Consortium provides charts listing all defined graphemes and their associated code points. In order to allow organizations to define their own private characters without conflicting with assigned Unicode characters, ranges of code points are left undefined. One of these ranges includes all of the code points between U+E000 and U+F8FF. Organizations may assign undefined code points to their own private graphemes.

One inherent problem with Unicode is that certain graphemes have been assigned to multiple code points. In an ideal system, each grapheme would be assigned to a single code point to simplify text processing. However, in order to encourage the adoption of the Unicode standard, character encodings such as ASCII were supported in Unicode. This resulted in certain graphemes being assigned to more than one code point in the Unicode standard.

Теоретические сведения

Префиксный код Хаффмана

Пусть нам дано сообщение **аааббвбддееаабееедгбав.**

Чтобы узнать наиболее выгодный префиксный код для такого сообщения, надо узнать частоту появления каждого символа в сообщении.

Подсчитайте и внесите в таблицу частоту появления каждого символа в сообщении.

Должно получиться:

а	б	в	г	д	е
6	4	2	1	3	5

Располагаем буквы в порядке возрастания их частоты.

1	2	3	4	5	6
г	в	д	б	е	а

Теперь возьмем два символа с наименьшей частотой и представим их листьями в дереве, частота которого будет равна сумме частот этих листьев.

Символы **г** и **в** превращаются в ветку дерева:

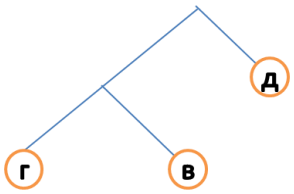
3	3	4	5	6
гв	д	б	е	а

Проделываем это до тех пор, пока не получится дерево, содержащее все символы.

Итак, сортируем таблицу:

Объединяем символ **д** и символ **гв** в ветку дерева:

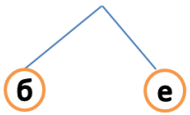
6	4	5	6
гвд	б	е	а



Сортируем:

4	5	6	6
б	е	гвд	а

9	6	6
бе	гвд	а



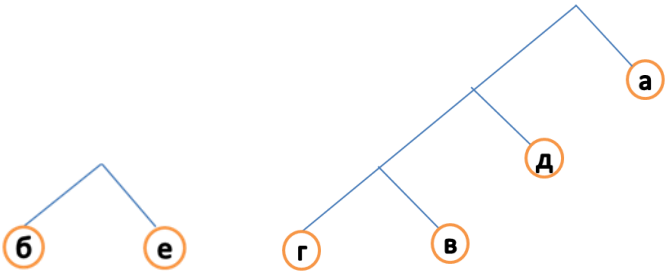
Сортируем:

6	6	9
гвд	а	бе

12	9
агвд	бе

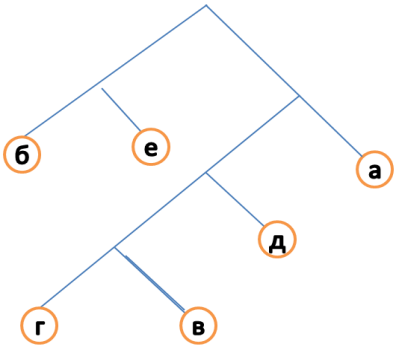
Сортируем:

9	12
бе	адгв



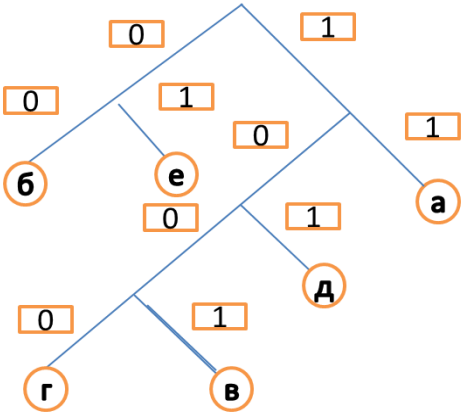
Получился префиксный код.

21
беадгв



Теперь осталось расставить 1 и 0. Пусть каждая правая ветвь обозначает 1, а левая – 0.

Составляем код буквы, идя по ветке дерева от буквы к основанию дерева. Тогда код для каждой буквы будет:



Тогда код для каждой буквы будет: а – 11, б – 00, е – 01, д – 101, г – 1000, в – 1001.

Задание 4

Построить префиксный код Хаффмена для сообщения «мамамыларамумыла».

Самостоятельно: постройте префиксный код Хаффмена для сообщения:

1) эникиэниэникэн;
2) бенибкибенибклё;
3) тертепертерапето;
4) кукаркккурекувек;
5) метротрометраам;
6) ленлентатамалён;
7) деньденометрдер;
8) вебервебнеровер;
9) диноинозавродин;
10) препипренаниепе;
11) тратрампатрмпик;
12) искискусствокусс.

Лабораторная работа 11
Кодирование графической, звуковой и видео информации

Теоретические сведения

1. Кодирование звуковой информации.

При оцифровке звука в памяти запоминаются только отдельные значения сигнала. Чем чаще записывается сигнал, тем лучше качество записи.

Частота дискретизации f – это количество раз в секунду, которое проходит преобразование аналогового звукового сигнала в цифровой. Измеряется в Герцах (Гц).

Глубина кодирования (а также, разрешение) – это количество бит, выделяемое на одно преобразование сигнала. Измеряется в битах (Бит).

Возможна запись нескольких каналов: одного (моно), двух (стерео), четырех (квадро).

Обозначим частоту дискретизации – f (Гц), глубину кодирования – B(бит), количество каналов – k, время записи – t (сек).

Количество уровней дискретизации d можно рассчитать по формуле:

$$d = 2^B.$$

Тогда объем записанного файла **V(бит) = f * B * k * t.**

Или, если дано количество уровней дискретизации,

$V(\text{бит}) = f * \log_2 d * k * t.$

Задание 1. Пройти по ссылке, изучить материал.

https://spravochnick.ru/informatika/kodirovanie_informacii/kodirovanie_zvukovoy_informacii/

Ответить на вопросы:

1. Что собой представляют непрерывные звуковые сигналы?

2. Чем отличаются дискретные звуковые сигналы от непрерывных?

3. Что собой представляет частота дискретизации звука?

4. Как вычислить информационный объем звукового файла?

5. Перечислить основные методы кодирования звуковой информации.

6. Привести примеры форматов звуковых файлов.

Пример 1

Производится двухканальная (стерео) звукозапись с частотой дискретизации 16 кГц и глубиной кодирования

32 бит. Запись длится 12 минут, ее результаты записываются в файл, сжатие данных не производится. Чему равен размер полученного файла (выразить в мегабайтах)?

$$V(\text{бит}) = f(\text{Гц})* B(\text{бит}) * k * t(\text{Сек}),$$

где V – размер файла, f – частота дискретизации, B – глубина кодирования, k – количество каналов, t – время.

Значит, $V(\text{Мб}) = (f * B * k * t) / 2^{23}$ (почему 2^{23}?).

Переведем все величины в требуемые единицы измерения:

$$V(\text{Мб}) = (16*1000 * 32 * 2 * 12 * 60) / 2^{23}$$

Представим все возможные числа, как степени двойки:

$V(\text{Мб}) = (2^4 * 2^3 * 125 * 2^5 * 2 * 2^2 * 3 * 15 * 2^2) / 2^{23} = (5625 * 2^{17}) / 2^{23} = 5625 / 2^6 = 5625 / 64 \approx 90$.

Задание 2

Привести свой пример вычисления размера полученного файла (см. пример 1).

Задание 3

Аналоговый звуковой сигнал был записан сначала с использованием 64 уровней дискретизации сигнала, а затем с использованием 4096 уровней дискретизации сигнала. Во сколько раз увеличился информационный объем оцифрованного звука?

$V(\text{бит}) = f * \log_2 d * k * t$, где V – размер файла, f – частота дискретизации, d – количество уровней дискретизации, k – количество каналов, t – время.

Теоретические сведения

2. Кодирование графической информации.

При оцифровке графического изображения качество картинки зависит от количества точек и количества цветов, в которые можно раскрасить точку.

Если X – количество точек по горизонтали,

Y – количество точек по вертикали,

I – глубина цвета (количество бит, отводимых для кодирования одной точки), то количество различных цветов в палитре $N = 2^I$. Соответственно, $I = \log_2 N$.

Тогда объем файла, содержащего изображение, V(бит) = X * Y * I

Или, если нам дано количество цветов в палитре, $V(\text{бит}) = X * Y * \log_2 N$.

Скорость передачи информации по каналу связи (пропускная способность канала) вычисляется как количество информации в битах, переданное за 1 секунду (бит/с).

Объем переданной информации вычисляется по формуле V = q * t, где q – пропускная способность канала, а t – время передачи.

Пример 2

Какой минимальный объём памяти (в Кбайт) нужно зарезервировать, чтобы можно было сохранить любое растровое изображение размером 64×64 пикселей при условии, что в изображении могут использоваться 256 различных цветов?

$V(\text{бит}) = X * Y * \log_2 N$, где V – объем памяти; X, Y – количество пикселей по горизонтали и вертикали, N – количество цветов.

$V\ (\text{Кб}) = (64 * 64 * \log_2 256) / 2^{13} = 2^{12} * 8 / 2^{13} = 4$

Задание 4

Для хранения растрового изображения размером 64х32 пикселя отвели 1 килобайт памяти. Каково максимально возможное число цветов в палитре изображения?

$V(бит) = X * Y * \log_2 N$, где V – объем памяти, X, Y – количество пикселей по горизонтали и вертикали, N – количество цветов.

Задание 5

Привести свой пример вычисления.

Теоретические сведения

3. Кодирование видеоинформации.

VirtualDub – свободная утилита для захвата, монтажа и редактирования видеопотока для платформы Windows, лицензированная на условиях GNU General Public License. Разработчик – Эвери Ли. Может использоваться для простых линейных операций над файлами формата видео.

Задание 6. Пройти по ссылке, изучить материал.
https://docs.google.com/presentation/d/1WAN4W7Peie1gzOyXvlLTJa0WSDpq9W5iVdZqT1FtXlk/htmlpresent

Ответить на вопросы:

1. Что включает в себя видеоинформация?
2. Как представляется видеоинформация?
3. Что собой представляет формат AVI?
4. Что такое «CODEC»?
5. Перечислить этапы обработки видеоинформация.
6. Что собой представляет оцифровка?
7. В чем состоит монтаж видеоклипа?

Задание 7

Скачать VirtualDub. Изучить возможности.

Лабораторная работа 12
Кодирование информации. Метод Шеннона–Фано

Теоретические сведения

Теоретический материал находится на портале университета (https://cspu.sharepoint.com/Education/Shared%20Documents/Forms/AllItems.aspx?viewid=c4d1b0a4%2D2b55%2D4e5b%2D9042%2D6d7bb4925514&id=%2FEducation%2FShared%20Documents%2FIITiMOI%2FISiT%2F2k%202020%2FTI%5FD%5FKn%2FTheory).

Примеры

1. Закодируйте фразу «Тише, мыши, тише, кот на крыше», используя метод Шеннона-Фано.

Задание. Закодировать фразу «наша саша шла по шоссе». Подсчитать количество бит, которое занимает закодированная фраза.

Задание для самостоятельной работы

Закодировать фразу «Сшит колпак да не по-колпаковски, надо колпак переколпаковать».

Таблица 3.1 – Пример кодирования методом Шеннона-Фано

Символ	**Кол-во**	**1-я цифра**	**2-я цифра**	**3-я цифра**	**4-я цифра**	**5-я цифра**	**Код**	**Кол-во бит**
пробел	5	0	0	0			000	15
ш	4	0	0	1			001	12
е	3	0	1	0			010	9
,	3	0	1	1			011	9
и	3	1	0	0			100	9
т	3	1	0	1	0		1010	12
ы	2	1	0	1	1		1011	8
к	2	1	1	1	0		1110	8
н	1	1	1	1	1		1111	4
о	1	1	1	0	0	0	11000	5
а	1	1	1	0	0	1	11001	5
м	1	1	1	0	1	0	11010	5
р	1	1	1	0	1	1	11011	5
								106

Незакодированная фраза – 30*8 бит = 240 бит.

Закодированная фраза – 106 бит.

Лабораторная работа 13

Кодирование информации. Метод Хаффмена

Теоретические сведения

Теоретический материал находится на портале университета (https://cspu.sharepoint.com/Education/Shared%20Documents/Forms/AllItems.aspx?viewid=c4d1b0a4%2D2b55%2D4e5b%2D9042%2D6d7bb4925514&id=%2FEducation%2FShared%20Documents%2FIITiMOI%2FISiT%2F2k%202020%2FTI%5FD%5FKn%2FTheory).

Примеры

1. Закодируйте фразу «Тише, мыши, тише, кот на крыше», используя метод Хаффмена.

Таблица 3.2 – Кодирование методом Хаффмена

Символ	**Кол-во**	11	10	9	8	7	6	5	4	3	2	1
пробел	5	5	5	5	5	5	6	6	8	10	12	18
ш	4	4	4	4	4	5	5	6	6	8	**10**	12
е	3	3	3	3	4	4	5	5	6	**6**	**8**	
,	3	3	3	3	3	4	4	5	**5**	**6**		
и	3	3	3	3	3	3	4	**4**	**5**			
т	3	3	3	3	3	3	**3**	**4**				
ы	2	2	2	3	3	**3**	**3**					
к	2	2	2	2	**3**	**3**						
н	1	2	2	**2**	**2**							
о	1	1	**2**	**2**								
а	1	**1**	**1**									
м	**1**	**1**										
р	**1**											

Окончание таблицы 3.2

1	2	3	4	5	6	7	8	9	10	11	**Код**	**Кол-во бит**
0	**1**	**00**	**01**	10	**10**	000	000	000	000	000	000	15
1	00	01	10	**11**	000	**001**	010	010	010	010	010	12
	01	10	11	000	001	010	**011**	110	110	110	110	9
		11	000	001	010	011	110	111	111	111	111	9
			001	010	011	110	111	100	100	100	100	9
				011	110	111	100	101	101	101	101	9
					111	100	101	**0010**	0011	0011	0011	8
						101	0010	0011	0110	0110	0110	8
							0011	0110	0111	**0111**	00101	5
								0111	**00100**	00101	001000	6
									00101	001000	001001	6
										001001	01110	5
											01111	5
												106

Незакодированная фраза – 30*8 бит = 240 бит.

Закодированная фраза – 106 бит.

Задание. Закодировать фразу «во дворе трава на траве дрова». Подсчитать количество бит, которое занимает закодированная фраза.

Задание для самостоятельной работы

Закодировать фразу «У Феофана Митрофаныча три сына – Феофанычи».

Индивидуальные работы 3–4

Кодирование методами Шеннона–Фано и Хаффмена

Закодировать фразу методами Шеннона-Фано и Хаффмена в соответствии с вариантом.

1. Ана, дэус, рики, паки,
Дормы кормы констунтаки,
Дэус дэус канадэус – бац!

2. One, two, Freddy's coming for you
Three, four, better lock your door
Five, six, grab a crucifix
Seven, eight, gonna stay up late.

3. Прибавь к ослиной голове
Еще одну, получишь две.
Но сколько б ни было ослов,
Они и двух не свяжут слов.

4. Эне-бене, рики-таки,
Буль-буль-буль,
Караки-шмаки
Эус-деус-краснодеус – бац!

5. Кони-кони, кони-кони,
Мы сидели на балконе,
Чай пили, воду пили,
По-турецки говорили.

11. Раз, два – упала гора;
три, четыре – прицепило;
пять, шесть – бьют шерсть;
семь, восемь – сено косим.

12. Плыл по морю чемодан,
В чемодане был диван,
На диване ехал слон.
Кто не верит – выйди вон!

13. Перводан, другодан,
На колоде барабан;
Свистель, коростель,
Пятерка, шестерка, утюг.

14. Мой котёнок очень странный,
Он не хочет есть сметану,
К молоку не прикасался
И от рыбки отказался.

15. Самолёт-вертолёт!
Посади меня в полёт!
А в полёте пусто –
Выросла капуста.

6. По-турецки говорили.
Чяби, чяряби
Чяряби, чяби-чяби.
Мы набрали в рот воды.

7. Тише, мыши, кот на крыше,
А котята ещё выше.
Кот пошёл за молоком,
А котята кувырком.

8. Эни-бени рити-Фати.
Дорба, дорба сентибрати.
Дэл. Дэл. Кошка. Дэл. Фати!

9. Кот пошёл за молоком,
А котята кувырком.
Кот пришёл без молока,
А котята ха-ха-ха.

10. Эне, бене, лики, паки,
Цуль, буль-буль,
Калики-цваки,
Эус-беус, клик-мадеус, бокс...

16. Цветом мой зайчишка – белый,
А ещё, он очень смелый!
Не боится он лисицы,
Льва он тоже не боится.

17. Ана-дэус-рики-паки,
Дормы-кормы-консту-таки,
Энус-дэус-кана-дэус-БАЦ!

18. Зуба зуба, зуба зуба,
Зуба дони дони мэ,
А шарли буба раз два три,
А ми раз два три замри.

19. Дрынцы-брынцы-бубен-цы,
Раз-звонились-удальцы,
Диги-диги-диги-дон,
Выхо-ди-скорее-вон!

20. Эни бэни рики паки
Турбаурбасентибряки.
Может – выйдет, может – нет,
В общем – полный Интернет

Лабораторная работа 14
Сжатие данных по методу Лемпеля–Зива

Теоретические сведения

Лемель и Зив используют следующую идею: если в тексте сообщения появляется последовательность из двух ранее уже встречавшихся символов, то эта последовательность объявляется новым символом, для нее назначается

код, который при определенных условиях может быть значительно короче исходной последовательности. В дальнейшем в сжатом сообщении вместо исходной последовательности записывается назначенный код. При декодировании повторяются аналогичные действия, и потому становятся известными последовательности символов для каждого кода. Одна из алгоритмических реализаций этой идеи включает следующие операции. Первоначально каждому символу алфавита присваивается определенный код (коды – порядковые номера, начиная с 0). При *кодировании*:

1. Выбирается первый символ сообщения и заменяется на его код.

2. Выбираются следующие два символа и заменяются своими кодами. Одновременно этой комбинации двух символов присваивается свой код. Обычно это номер, равный числу уже использованных кодов. Так, если алфавит включает 8 символов, имеющих коды от 000 до 111, то первая двухсимвольная комбинация получит код 1000, следующая – код 1001 и т.д.

3. Выбираются из исходного текста очередные 2, 3, ..., N символов до тех пор, пока не образуется еще не встречавшаяся комбинация. Тогда этой комбинации присваивается очередной код, и поскольку совокупность A из первых N – 1 символов уже встречалась, то она имеет свой код, который и записывается вместо этих N – 1 символов. Каждый акт введения нового кода назовем шагом кодирования.

4. Процесс продолжается до исчерпания исходного текста. При *декодировании* код первого символа, а затем второго и третьего символов, заменяются на символы алфавита. При этом становится известным код комбинации

второго и третьего символов. В следующей позиции могут быть только коды уже известных символов и их комбинаций. Процесс декодирования продолжается до исчерпания сжатого текста.

Сколько двоичных разрядов нужно выделять для кодирования? Ответ может быть следующим: число разрядов R на каждом шаге кодирования равно числу разрядов в наиболее длинном из использованных кодов (т.е. числу разрядов в последнем использованном порядковом номере). Поэтому если последний использованный код (порядковый номер) равен $(13)_{10} = (1101)_2$, то коды A всех комбинаций должны быть четырехразрядными при кодировании вплоть до появления номера 16, после чего все коды символов начинают рассматриваться как пятиразрядные (R = 5).

Пример. Пусть исходный текст представляет собой двоичный код (первая строка таблицы 6), т.е. символами алфавита являются 0 и 1. Коды этих символов соответственно также 0 и 1. Образующийся по методу Лемпеля-Зива код (LZ-код) показан во второй строке таблицы 3.3. В третьей строке отмечены шаги кодирования, после которых происходит переход на представление кодов A увеличенным числом разрядов R. Так, на первом шаге вводится код 10 для комбинации 00, и поэтому на следующих двух шагах R = 2, после третьего шага R = 3, после седьмого шага R = 4, т.е. в общем случае R = K после шага $2^{K-1} – 1$.

Таблица 3.3 – Пример кодирования методом Лемпеля-Зива

Исходный текст	0	00	000	01	11	111	1111	110	0000	00000	1101	11011
LZ-код	0	00	100	001	011	1011	1101	1010	0000	10010	1101	10111
R		2		3				4				
Вводимые коды	–	10	11	100	101	110	111	1000	1001	1010	1011	1100

В приведенном примере LZ-код оказался длиннее исходного кода, так как обычно короткие тексты не дают эффекта сжатия. Эффект сжатия проявляется в достаточно длинных текстах и особенно заметен в графических файлах.

В другой известной реализации LZ-метода любая ранее встречавшаяся последовательность в сжатом тексте представляет собой совокупность следующих данных:

– номер первого символа в ранее встречавшейся последовательности;

– число символов в последовательности;

– следующий символ в текущей позиции кодируемого текста.

Алгоритм LZ78

Алгоритм LZ78 вносит все встреченные им последовательности в словарь. Всякий раз, когда среди данных, которые надо сжать, встречается последовательность, программа обращается к словарю:

– если последовательность находится в словаре, то в выходной файл заносится код для этой записи;

– если последовательность представляет собой расширенный вариант последовательности из словаря, то она добавляется в таблицу.

Таблица 3.4 – Пример кодирования методом LZ78

Исходный символ	Словарь	Код
a	–	–
ab	257	a
bc	258	b
ca	259	c
ab (257)		
abc	260	257
ca (259)		
cab	261	259
bc (258)		258

Пример. Рассмотреть, как программа LZ78 читает каждый байт abcabcabc (см. табл. 3.4).

Получен код: abc 257 259 258.

При раскодировании получаем: abc ab ca bc.

Задание 1. Сжать следующие данные, используя двухсимвольный алфавит, имеющий коды от 0 до 1, 1 01 10 101 1010 011. Выполнить обратную операцию.

Задание 2. Сжать последовательность 1) abcdabcd, 2) abaabcabcdabcd, используя алгоритм LZ78. Привести пример распаковки.

Самостоятельно:

a) используйте 4-символьный алфавит (00, 01, 10, 11) для сжатия данных;

b) сожмите последовательность «всемвсемвсемикаждомускажу».

Индивидуальная работа 5

Сжатие данных по методу Лемпеля–Зива

1. Используя двухсимвольный алфавит (0, 1) закодировать следующую фразу:

Вариант 1	Вариант 2
0100101010010000101	0001000010101001101
Вариант 3	Вариант 4
1110100110111001101	0001000010101001101
Вариант 5	Вариант 6
0100101010010000101	10110111100110001101
Вариант 7	Вариант 8
0010100110010000001	01011011001010101011
Вариант 9	Вариант 10
10101001101100111010	000100101100100010001
Вариант 11	Вариант 12
01011011010100010001	11010101100110000 1001
Вариант 13	Вариант 14
0001011101101001 11	1010001001010100 01011
Вариант 15	Вариант 16
0100001000000100001	0100101010010000101
Вариант 17	Вариант 18
0100100010010000101	0001010010101001101
Вариант 19	Вариант 20
1110100110110001101	0001001010101001101

2. Закодировать следующую фразу, используя код LZ78:

Вариант 1	Вариант 2
кукуркукурекурекун	упупапекапекаупуп
Вариант 3	Вариант 4
лорлоралоранранлоран	пропронепронепрне-пронас

Вариант 5
какатанекатанекатата
Вариант 6
менменаменаменатеп
Вариант 7
долделдолдилделдил
Вариант 8
sarsalsarsanlasanl
Вариант 9
kloklonkolonklonkl
Вариант 10
tertrektekertektrek
Вариант 11
bigbonebigborebigbo
Вариант 12
commercommecommerce
Вариант 13
webwerbweberweberweb
Вариант 14
porpoterpoterporter
Вариант 15
mantopmentopomantomen
Вариант 16
roporopoterropoterter
Вариант 17
webwerbweberweberweb
Вариант 18
sionsinossionsinos
Вариант 19
comconcomconacom
Вариант 20
mantopmentopomantomen

Вопросы к модулю 3

1. Кем заложены теоретические основы сжатия информации?

2. Что лежит в основе алгоритма Лемпеля–Зива.

3. В чем заключается метод повторяющихся последовательностей?

4. В основе какого метода лежит идея замены часто встречающихся последовательностей символов в файле ссылками на образцы?

5. Какая система счисления должна использоваться в компьютере с целью минимизации количества элементов в устройствах хранения?

6. Какая схема кодирования называется разделимой?

7. Является ли префиксная схема кодирования разделимой?

8. Какие методы относятся к вероятностным методам?

9. Перечислить единицы измерения пропускной способности канала, энтропии, скорости.

10. Какие помехи вызывают внешние источники помех?

11. Дана схема кодирования $\langle \delta = a \to 01, b \to 110, c \to 101 \rangle$. Удовлетворяет ли данная схема неравенству МакМиллана?

12. Является ли разделимая схема префиксной?

13. Задача оптимального кодирования. Метод Шеннона–Фано.

14. Суть метода Хаффмена.

15. Перечислить основные технические характеристики процессов сжатия.

16. Какое сжатие называется обратимым сжатием? Необратимым сжатием?

17. Что приводит к снижению объема выходного потока информации без изменения его информативности?

МОДУЛЬ 4. ПОМЕХОУСТОЙЧИВОЕ КОДИРОВАНИЕ

Лабораторная работа 15

Коды Хемминга

Теоретические сведения

1. Обнаружение одиночной ошибки (1)

Наиболее известные из самоконтролирующихся и самокорректирующихся кодов – коды Хемминга. Построены они применительно к двоичной системе счисления.

Для построения самокорректирующегося кода достаточно иметь один контрольный разряд (код с проверкой на четность). Но при этом мы не получаем никаких указаний о том, в каком именно разряде произошла ошибка, и, следовательно, не имеем возможности исправить ее. Остаются незамеченными ошибки, возникшие в четном числе разрядов.

Коды Хемминга имеют большую относительную избыточность, чем коды с проверкой на четность, и предназначены либо для исправления одиночных ошибок (при d = 3), либо для исправления одиночных и обнаружения без исправления двойных ошибок (d = 4). В таком коде n-значное число имеет m информационных и k контрольных разрядов. Каждый из контрольных разрядов является знаком четности для определенной группы информационных знаков слова. При декодировании производится k групповых проверок на четность. В результате каждой проверки в соответствующий разряд регистра ошибки записывается 0, если проверка была успешной, или 1, если была обнаружена нечетность.

Группы для проверки образуются таким образом, что в регистре ошибки после окончания проверки получается K-разрядное двоичное число, показывающее номер

позиции ошибочного двоичного разряда. Изменение этого разряда – исправление ошибки.

Первоначально эти коды предложены Хеммингом в таком виде, при котором контрольные знаки занимают особые позиции: позиция i-го знака имеет номер 2^{i-1}. При этом каждый контрольный знак входит лишь в одну проверку на четность.

Рассмотрим код Хемминга, предназначенный для исправления одиночных ошибок, т.е. код с минимальным кодовым расстоянием d = 3.

Ошибка возможна и в одной из n позиций. Следовательно, число контрольных знаков, а значит, и число разрядов регистра ошибок должны удовлетворять условию

$$k \geq \log_2(n + 1). \quad (1)$$

Отсюда

$$m \leq n - \log_2(n + 1). \quad (2)$$

Значения m и k для некоторых коротких кодов, вычисленные по формулам (1) и (2) даны в табл. 8.

Таблица 4.1 – Значения m, n, k

n	3	4	5	6	7	8	9	10	11	12
m	1	1	2	3	4	4	5	6	7	8
k	2	3	3	3	3	4	4	4	4	4

Чтобы число в регистре ошибок (РОШ) указывало номер позиции ошибочного разряда, группы для проверки выбираются по правилу:

I гр.: *все нечетные позиции, включая и позиции контрольного разряда, т.е. позиции, в первом младшем разряде которых стоит 1.*

II гр.: *все позиции, номера которых в двоичном представлении имеют 1 во втором разряде справа (например, 2, 3, 6, 7, 10) и т. д.*

III гр.. *разряды, имеющие 1 в третьем разряде справа, и т. д.*

Примечание: каждый контрольный знак входит только в одну проверяемую группу.

Пример 1. Пусть k = 5 (табл. 9).

Таблица 4.2 – Формирование контрольных групп

Номер проверки	Позиция контрольного знака	Проверяемые позиции
1	*1*	*1, 3, 5, 7, 9, 11, 13, ...*
2	*2*	*2, 3, 6, 7, 10, 11, ...*
3	*4*	*4, 5, 6, 7, 12, 13, ...*
4	*8*	*8, 9, 10, 11, 12, 13, ...*
5	*16*	*16, 17, 18, 19, 20, 21*

Пример 2. Рассмотрим семизначный код Хемминга, служащий для изображения чисел от 0 до 9 (табл. 4.3).

Таблица 4.3 – Семизначный код Хемминга

Десятичное число	Простой двоичный код				Код Хемминга						
								к		к	к
0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	1	0	0	0	0	1	1	1
2	0	0	1	0	0	0	1	1	0	0	1
3	0	0	1	1	0	0	1	1	1	1	0
4	0	1	0	0	0	1	0	1	0	1	0
5	0	1	0	1	0	1	0	1	1	0	1
6	0	1	1	0	0	1	1	0	0	1	1
7	0	1	1	1	0	1	1	0	1	0	0
8	1	0	0	0	1	0	0	1	0	1	1
9	1	0	0	1	1	0	0	1	1	0	0

Пусть передан код числа 6 в виде «0 1 1 0 0 1 1», а приняли в виде «0 1 0 0 0 1 1». Проверочные группы:

I проверка: разряды 1, 3, 5, 7 – дает 1 в младший разряд РОШ.

II проверка: разряды 2, 3, 6, 7 – дает 0 во второй разряд РОШ.

III проверка: разряды 4, 5, 6, 7 – дает 1 в третий разряд РОШ.

Содержимое РОШ «101», значит ошибка в пятой позиции.

Примечание. В каждый из контрольных разрядов при построении кода Хемминга посылается такое значение, чтобы общее число единиц в его контрольной сумме было четным. РОШ заполняется, начиная с младшего разряда.

Рост кодового расстояния позволяет увеличить корректирующую способность кода. В то время как d = 2 у кода с проверкой на четность позволяет обнаруживать одиночную ошибку, код Хемминга с d = 3 исправляет ее.

2. Обнаружение одиночной ошибки (2)

В коде Хемминга вводится понятие кодового расстояния d (расстояния между двумя кодами), равного числу разрядов с неодинаковыми значениями. Возможности исправления ошибок связаны с минимальным кодовым расстоянием d_{min}. Исправляются ошибки кратности r = ent (d_{min} – 1) / 2 и обнаруживаются ошибки кратности d_{min} – 1 (здесь ent означает «целая часть»). Так, при контроле на нечетность d_{min} = 2 и обнаруживаются одиночные ошибки. В коде Хемминга d_{min} = 3. Дополнительно к информационным

разрядам вводится $L = \log_2 K$ избыточных контролирующих разрядов, где K – число информационных разрядов, L округляется до ближайшего большего целого значения. L-разрядный контролирующий код есть инвертированный результат поразрядного сложения (т.е. сложения по модулю 2) номеров тех информационных разрядов, значения которых равны 1.

Пример 1. Пусть имеем основной код 100110, т.е. K = 6. Следовательно, L = 3 и дополнительный код равен 010 # 011 # 110 = 111, где # – символ операции поразрядного сложения по модулю 2. После инвертирования имеем 000.

Вместе с основным кодом будет передан и дополнительный. На приемном конце вновь рассчитывается дополнительный код и сравнивается с переданным. Фиксируется код сравнения (поразрядная операция отрицания равнозначности), и если он отличен от нуля, то его значение есть номер ошибочно принятого разряда основного кода. Так, если принят код 100010, то рассчитанный в приемнике дополнительный код равен инверсии от 010 # 110 = 100, т.е. 011, что означает ошибку в 3-м разряде.

Пример 2. Основной код 1100000, дополнительный код 110 (результат инверсии кода 110 # 111 = 001). Пусть принятый код 1101000, его дополнительный код 010, код сравнения 100, т.е. ошибка в четвертом разряде.

Задание 1

1. Передан код числа 5 в виде «0 1 0 1 1 0 1», а приняли в виде «0 1 0 1 0 0 1». Обнаружить позицию ошибки.

2. Передан код числа 8 в виде «1 0 0 1 0 1 1», а приняли в виде «1 0 1 1 0 1 1». Обнаружить позицию ошибки.

Задание 2

1. Основной код 100110, принятый код 100100. Подсчитать дополнительный код. Обнаружить позицию ошибки.

2. Основной код 0101100, принятый код 0111100. Подсчитать дополнительный код. Обнаружить позицию ошибки.

Задание для самостоятельной работы

1. Передан код числа 7 в виде «0 1 1 0 1 0 0», а приняли в виде «1 1 1 0 1 0 0». Обнаружить позицию ошибки.

2. Основной код 0101100, принятый код 0101101. Подсчитать дополнительный код. Обнаружить позицию ошибки.

3. Познакомиться с программой Code (Папка «Учебник по помехоустойчивому кодированию» на сервере университета).

Лабораторная работа 16
Электронная цифровая подпись

Теоретические сведения

Шифр Эль-Гамаля

Описание метода. Для всей группы абонентов выбираются некоторое большое простое число p и число g, такие, что различные степени g суть различные числа по модулю p. Числа p и g передаются абонентам в открытом виде (они могут использоваться всеми абонентами сети). Затем каждый абонент группы выбирает свое секретное число c_i, $1<c_i<p-1$, и вычисляет соответствующее ему открытое число d_i,

$$d_i = g^{ci} \bmod p \,. \qquad (1)$$

В результате получаем таблицу:

Таблица 4.4 – Ключи пользователей в системе Эль-Гамаля

Абонент	Секретный ключ	Открытый ключ
A	c_A	d_A
B	c_B	d_B
C	c_C	d_C

Покажем теперь, как A передает сообщение m абоненту B. Будем предполагать, что сообщение представлено в виде числа $p_m < p$.

Шаг 1. A формирует случайное число k, $1 < k < p - 2$, вычисляет числа

$$r = g^k \bmod p \quad (2)$$

$$e = m \cdot d_B^k \bmod p \quad (3)$$

и передает пару чисел (r, e) абоненту B.

Шаг 2. B, получив (r,e), вычисляет

$$m' = e \cdot r^{p-1-C_B} \bmod p \quad (4)$$

Свойства шифра Эль-Гамаля. Абонент B получил сообщение, т.е. $m'=m$; противник, зная p, g, d_B, r и e, не может вычислить m.

Пример. Передадим сообщение m=17 от A к B. Выберем параметры: p=31, g=7. Пусть абонент B выбрал для себя секретное число $c_B=19$ и вычислил по (1)

$d_B=7^{19} \bmod 31 = 14$.

Абонент A выбирает случайно число k, например, k=11, и вычисляет по (2), (3):

$r=7^{11} \bmod 31 = 20$,

$e=17 \cdot 14^{11} \bmod 31 = 29$.

Теперь A посылает к B зашифрованное сообщение в виде пары чисел (20, 29). B вычисляет по (4)

$m'=29\cdot 20^{31-1-19} \bmod 31 = 29\cdot 20^{11} \bmod 31 = 17$.

Видим, что B смог расшифровать переданное сообщение.

Ясно, что по аналогичной схеме могут передавать сообщения все абоненты в сети. Заметим, что любой абонент, знающий открытый ключ абонента B, может посылать ему сообщения, зашифрованные с помощью открытого ключа d_B, но только абонент B, и никто другой, может расшифровать эти сообщения, используя известный только ему секретный ключ c_B.

Объем шифра в два раза превышает объем сообщения, но требуется только одна передача данных (при условии, что таблица с открытыми ключами заранее известна всем абонентам).

Задание 1. Передать сообщение m = 15 от A к C. Выберем параметры: p = 23, g = 5. Пусть абонент C выбрал для себя секретное число $c_c = 13$ и вычислил по (1).

Абонент A выбирает случайно число k, например, k = 7, и вычисляет по (2), (3). Теперь A посылает к C зашифрованное сообщение в виде пары чисел (r, e). C вычисляет по (4).

Задание 2. Передадим сообщение m от C к B. Выбрать параметры самостоятельно.

Алгоритм цифровой подписи DSA

Данный алгоритм является частью Американского стандарта DSS. В алгоритме используется однонаправленная хэш-функция H(x). Стандарт определяет использование алгоритма SHA-1.

Параметры:

p – простое число L битов, где L принимает значение кратное 64 в диапазоне от 512 до 1024;

q – 160 – битовый множитель p – 1;

$a = g^{(p-1)/q} \bmod p$, где $g < p - 1$, для которого $g^{(p-1)/q} \bmod p > 1$;

$y = a^x \bmod p$, где $x < q$;

m – текст.

Ключ подписания: y, p, q, a. Ключ верификации: x.

Подписание:

k – случайное число, $k < q$;

$r = (a^k \bmod p) \bmod q$ – вычисление первой части подписи;

$s = (k^{-1}(H(m) + xr)) \bmod q$ – вычисление второй части подписи.

Подпись: (r, s).

Верификация:

$$w = s^{-1} \bmod q;$$

$$u_1 = (H(m) \square w) \bmod q;$$

$$u_2 = (rw) \bmod q;$$

$$v = ((a^{u1} \cdot y^{u2}) \bmod p) \bmod q,$$

если v = r – то подпись подлинная.

Пример. Параметры домена q = 13, p = 4q + 1 = 53 и g = 16.

$a = g(p - 1)/q \bmod p = 164 \bmod 53 = 28$

Предположим, что ключевая пара имеет вид x = 3 и

$y = a^x \pmod p = 28^3 \pmod{53} = 10$.

Если мы хотим подписать сообщение с хэш-значением $H = 2$, то сначала нужно выбрать ключ $k = 3$ (должно быть взаимно-просто с p) и найти

$r = (a^k \pmod p) \pmod q = (28^3 \pmod{53}) \cdot \pmod{13} = 44 \bmod 13 = 10$,

$s = (H + xr)/k \pmod q = (2 + 3 \cdot 10) \cdot 3^{-1} \pmod{13} = (2 + 3 \cdot 10) \cdot 9 \pmod{13} = 2$.

Для проверки подписи получатель определяет

$A = H/s \pmod q = 2 \cdot 2^{-1} \pmod{13} = 2 \cdot 7 \pmod{13} = 1$,

(2^{-1} означает нахождение обратного к 2 по модулю 13).

Это число 7. 2 * 7 ≡ 1 mod 13; 14 ≡ 1 mod 13, то есть дает остаток 1 при целочисленном делении 40 на 13, 14 = 13 * 1 + 1)

$B = r/s \pmod q = 10 \cdot 2^{-1} = 10 \cdot 7 \pmod{13} = 5$,

$v = (g^A y^B \pmod p) \pmod q = (16^1 \cdot 10^5 \pmod{53}) \pmod{13} = (16 \cdot 100000 \pmod{53}) \pmod{13} = 36 \pmod{13} = 10$.

Ввиду равенства $v = r = 10$ делаем вывод о корректности подписи.

Задание 3. Подписать сообщение с хэш-значением H. Параметры домена: $p = 4q + 1$.

Таблица 4.5 – Варианты заданий для самостоятельной работы

№	q	g	x	k	H
1.	31	26	3	11	7
2.	19	26	11	7	3
3.	43	26	13	19	7
4.	53	26	19	23	3
5.	17	26	7	11	7
6.	37	26	23	13	3
7.	31	28	3	11	7
8.	19	28	11	7	3
9.	43	28	13	19	7
10.	53	28	19	23	3
11.	17	28	7	11	7
12.	37	28	23	13	3

Лабораторная работа № 17

Зашита кейса по помехоустойчивому кодированию

Индивидуальная работа 6

Разработка кейса по помехоустойчивому кодированию

Разработать кейс с описанием одного из помехоустойчивых кодов.

Таблица 4.6 – Задания для кейса

1. Инверсный	11. С постоянным весом
2. Голея	12. Корреляционный
3. БЧХ	13. С четным числом 1
4. Грея	14. Миласа–Абрамсона
5. Рида–Маллера	15. По максимуму правдоподобия
6. С пороговым декодированием	16. С последовательным декодированием
7. Мажоритарный	17. Варшамова
8. Файра	18. Сверточный
9. Абрамсона	19. Плоткина
10. Рида–Соломона	20. Итеративный

Вопросы к модулю 4

1. Как осуществляется передача дискретных сообщений по каналу с шумами.
2. В чем смысл помехоустойчивого кодирования?
3. Какие существуют помехоустойчивые коды?
4. Дать определение избыточности.
5. Перечислить виды избыточности.
6. Какие помехи вызывают внешние источники помех?

7. Что понимают под кодовым расстоянием?

8. Чему равно кодовое расстояние между двумя кодовыми комбинациями (например, 011 и 101)?

9. Какие искажения называются краевыми? Дробления?

10. Какие коды называются непрерывными? Разделимыми?

11. В чем смысл кода Хемминга?

12. Какой код является оптимальным?

13. Что применяется для обнаружения и исправления ошибок в сотовых системах связи (ССС)?

14. Дать определение понятиям интерливинг, адаптивная коррекция, Antenna Diversity.

15. В чем причина затухания сигнала в ССС?

16. Какие существуют негативные факторы при передаче сигналов?

17. В чем состоит преимущество передачи цифровых сигналов перед аналоговыми?

18. Какие коды применяются в ССС для обнаружения ошибок?

19. Какие коды применяются в ССС для исправления ошибок?

20. Что представляет собой цифровая подпись?

21. Для чего используется секретный ключ? Открытый ключ?

22. Как проверить корректность цифровой подписи?

23. Как работает алгоритм Эль-Гамаля?

24. В чем состоит особенность алгоритма RSA? DSA? Шнорра?

25. Что понимают под стойкостью криптографического алгоритма?

ВОПРОСЫ ДЛЯ ИТОГОВОГО ТЕСТИРОВАНИЯ

1. Среди перечисленных определений одно не относится к определению понятия «информация»:

1) сообщение о состоянии и свойствах объекта, явления, процесса;

2) последовательность действий, направленных на достижение цели;

3) содержание сигналов, поступающих в кибернетическую систему из окружающей среды, которое может быть использовано для целей управления системой.

2. Для преобразования сигналов из непрерывной формы в дискретную используются преобразователи:

1) код-аналог;

2) аналого-цифровые;

3) цифро-аналоговые.

3. Любой непрерывный сигнал, имеющий ограниченный спектр частот, полностью определяется последовательностью своих мгновенных значений, отсчитанных через интервалы времени $\Delta t = \frac{1}{2 f_c}$, где f_c – верхняя граничная частота спектра непрерывного сигнала. Это теорема:

1) А. Хемминга;

2) К. Шеннона;

3) В. Котельникова.

4. Определять информационную емкость системы предложил:

1) К. Шеннон;

2) А. Харкевич;

3) Р. Хартли;

4) Ю. Шрейдер.

5. Полезность или ценность информации для выполнения функций управления предложил определять:

1) К. Шеннон;

2) Ю. Шрейдер;

3) Р. Хартли;

4) А. Харкевич.

6. Пусть алфавит содержит четыре символа и их вероятности равны соответственно $p_1 = p_2 = p_3 = p_4 = 0{,}25$. Чему будет равна неопределенность (энтропия) H?

1) 1;

2) 2;

3) 4.

7. Единица измерения степени неопределенности, содержащаяся в одном опыте, имеющем два равновероятных исхода, называется:

1) бит;

2) байт;

3) бит/буква;

4) бит/сек.

8. Чему будет равно количество информации, если событие имеет два равновероятных исхода:

1) ½;

2) 1;

3) 2.

9. Поставить в соответствие подходу его характеристику:

1) структурный;

2) статистический;

3) прагматический;

4) семантический.

a) учет вероятностных характеристик источника сообщений;

b) оценка сообщения с точки зрения получения лучшего управленческого решения;

c) опирается на тезаурусный подход;

d) определение меры количества информации;

e) определение информационной емкости сообщения.

10. Опыт γ состоит в выполнении опытов α и β. Поставить в соответствие:

1) H_{γ};

2) H_{α};

3) $I_{\alpha,\beta}$.

a) условная энтропия;

b) безусловная энтропия;

c) энтропия сложного опыта;

d) количество информации.

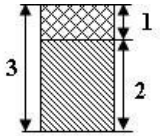
11. Если опыт заключается в одновременной реализации опытов α и β, независимых друг от друга, то энтропия:

1) $H_{\gamma} > H_{\alpha} + H_{\beta}$;

2) $H_{\gamma} = H_{\alpha} + H_{\beta}$;

3) $\cdot H_{\gamma} \le H_{\alpha} + H_{\beta}$

12. На рисунке показана зависимость энтропии и количества информации.



Поставить в соответствие:

1) 1;
2) 2;
3) 3.

a) H_{β};
b) $H_{\beta/\alpha}$;
c) $H_{\beta/\beta}$;
d) $I_{\alpha,\beta}$.

13. Наибольшую неопределенность среди всех опытов, имеющих n исходов, имеет опыт, у которого исходы:

1) зависят от предыдущего опыта;
2) неравновероятны;
3) равновероятны.

14. Количество какой информации может быть вычислено по формуле

$$I = \log \frac{p_1}{p_0},$$

если p_1 – вероятность достижения цели после получения информации о событии, p_0 – вероятность достижения цели до получения информации о событии?

1) семантической;
2) синтаксической;
3) прагматической.

15. Восстановить правильную последовательность системы связи:

1) канал связи;
2) получатель сообщения;
3) источник сообщения;
4) передатчик;
5) приемник.

16. В избыточных кодах имеет место соотношение:

1) $d = r + s - 1$, $r \geq s$;

2) $d = 1 + r + s$, $r \geq s$;

3) $d = 1 + r + s$, $r \leq s$.

17. Кодовое расстояние d = 2. Выбрать свойство кода.

1) обнаруживает одну ошибку;

2) обнаруживает две ошибки;

3) отличает одну кодовую комбинацию от другой.

18. Если из $\beta_{i1}\beta_{i2}...\beta_{ik} = \beta_{j1}\beta_{j2}...\beta_{jl}$ следует, что $k = l$, $\forall t \in \{1,2,...,k\}$ и $i_t = j_t$, то схема кодирования называется:

1) префиксной;

2) постфиксной;

3) разделимой.

19. Префиксная схема кодирования является разделимой?

1) да;

2) нет.

20. Пусть A = {a, b}, B = {0, 1} и $\delta = \langle a \to 010, b \to 01 \rangle$ $l_\delta(P) = \sum_{i=1}^{n} P_i l_i$.

Чему будет равна средняя длина кодирования δ, если $P = \langle 0.2, 0.8 \rangle$?

1) 1;

2) 1.4;

3) 2;

4) 2.2.

21. Алгоритм Фано строит схему кодирования:

1) оптимальную префиксную;

2) разделимую префиксную;

3) оптимальную постфиксную.

22. Распределение вероятностей и код Хаффмена представлены в таблице:

P_i	код
0.1	10
0.2	11
0.18	00
0.17	001
0.15	010
0.15	0110
0.01	0111

Чему равно $l_\delta(P)$?

1) 2.56;
2) 2.76;
3) 2.94;
4) 3.

23. Поставить в соответствие.

1) C бит/с;
2) H бит/букв;
3) V букв/с.

a) пропускная способность канала;
b) энтропия источника информации;
c) скорость передачи букв по каналу;
d) средняя длина кода сообщения.

24. Передача букв по каналу со скоростью $V < \frac{C}{H}$ возможна?

1) да;
2) нет.

25. Пусть вероятности появления букв $\alpha_1, \alpha_2, \alpha_3, \alpha_4$

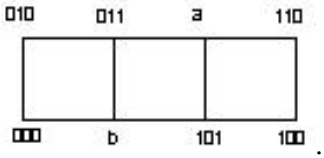
в сообщении равны $p_1 = 0{,}125$, $p_2 = 0{,}5$, $p_3 = 0{,}125$, $p_4 = 0.25$, C = 10000 бит/с. Чему равна максимальная скорость передачи сообщения?

1) 2500 букв/с;
2) 5000 букв/с;
3) 5714 букв/с.

26. Число разрядов, на которые отличаются любые две кодовые комбинации, называется:

1) оптимальным кодом;
2) коэффициентом сжатия;
3) кодовым расстоянием.

27. Трехэлементный двоичный код представлен в виде графа:



Чему равны a и b?

1) a – 011, b – 001;
2) a – 111, b – 001;
3) a – 001, b – 111.

28. Чему равно кодовое расстояние между двумя кодовыми комбинациями 111 и 100?

1) 1;
2) 2;
3) 3.

29. Пусть A = {c, d}, B = {0, 1} и $\delta = \langle c \to 01, d \to 0 \rangle$. Схема δ:

1) разделима;

2) неразделимая;

3) префиксная.

30. Дана схема $\delta = \langle a \to 01, b \to 110, c \to 101 \rangle$. Удовлетворяет ли данная схема неравенству МакМиллана?

1) да;

2) нет.

31. Дана схема $\delta = \langle a \to 01, b \to 10, c \to 101 \rangle$. Чему равно значение $\sum_{i=1}^{n} \left(\frac{1}{2} \right)^{l_i}$?

1) ½;

2) 3/8;

3) 5/8;

4) 1.

32. Является ли префиксная схема неразделимой?

1) да;

2) нет.

ЗАКЛЮЧЕНИЕ

В настоящее время уделяется большое внимание вопросам фундаментализации обучения информатике. В учебных планах вузов предусмотрено изучение теоретической информатики, как по направлениям подготовки бакалавров, так и по направлениям подготовки магистров.

В практикуме представлены материалы для изучения курса «Теория информации, данные, знания».

Для организации самостоятельной работы студентов, обучающихся по направлению «Информационные системы и технологии», в пособии представлены индивидуальные задания и вопросы к тестам.

Пособие может быть использовано при обучении по программам бакалавриата и магистратуры. При изучении дисциплины «Теория информации, данные, знания» используется модульно-рейтинговая система контроля знаний студентов. В приложении приведен пример оценки сформированности компетенций по указанной дисциплине.

ВОПРОСЫ К ЭКЗАМЕНУ

1. Понятие и формы представления информации, свойства информации.

2. Объем сигнала.

3. Емкость канала связи.

4. Данные. Особенности (свойства) декларативных знаний.

5. Знания. Классификация знаний.

6. Формы представления информации.

7. Измерение информации.

8. Энтропия как мера степени неопределенности.

9. Энтропия сложных событий.

10. Моделирование данных.

11. Количество информации.

12. Семантический подход к определению количества информации.

13. Прагматический подход к определению количества информации.

14. Формула Р. Хартли.

15. Алгоритмическое измерение количества информации.

16. Концепция разнообразия Эшби.

17. Информация как мера неоднородности по Глушкову.

18. Кодирование информации, алфавиты, системы счисления.

19. Избыточность.

20. Оптимальное кодирование. Метод Шеннона–Фано.

21. Передача дискретных сообщений по каналу с шумами.

22. Помехоустойчивое кодирование.

23. Алфавитное кодирование.

24. Корректирующие коды. Коды Хемминга.

25. Алгоритмы сжатия данных.

26. Метод повторяющихся последовательностей.

27. Алгоритм Лемпеля–Зива.

28. Криптосистема Эль-Гамаля.

29. Электронная цифровая подпись.

30. Виды помех и борьба с ними.

31. Стандарты сотовой связи.

32. Интерливинг. Перемежение.

33. Адаптивная коррекция.

34. Понятие информационной системы.

35. Образовательные ИС.

36. Геоинформационные системы.

37. Интеллектуальные информационные системы.

38. Экспертные системы в образовании.

39. Мировые информационные ресурсы. Определение.

40. Государственные (национальные) информационные ресурсы.

41. Информационные ресурсы предприятий.

42. Персональные информационные ресурсы.

43. Глобальные информационные сети.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Безручко, В.Т. Информатика: (курс лекций): учеб. пособие для вузов / В.Т. Безручко. – Москва: ИНФРА-М, 2018. – ISBN 978-5-8199-0763-4.
2. Велихов, А.В. Основы информатики и компьютерной техники: учеб. пособие для вузов / А. Велихов. – Москва: СОЛОН Пресс, 2007. – ISBN 5-98003-022-0.
3. Гагарина, Л.Г. Современные проблемы информатики и вычислительной техники: учеб. пособие для подготовки магистров / Л.Г. Гагарина, А.А. Петров. – Москва: Инфра-М, 2013. – ISBN 978-5-8199-0442-8.
4. Информатика: базовый курс: учеб. пособие для техн. вузов / ред. С.В. Симонович. – 3-е изд. – Санкт-Петербург: Питер, 2012. – ISBN 978-5-4461-0842-8.
5. Информатика и информационные технологии: учеб. пособие для вузов / ред. Ю.Д. Романова. – 4-е изд., перераб. и доп. – Москва: Эксмо, 2010. – ISBN 978-5-699-35357-6.
6. Колмогоров, А.Н. Три подхода к определению понятия «Количество информации» / А.Н. Колмогоров // Новое в жизни, науке, технике. – Сер. «Математика, кибернетика». – № 1. – 1991. – С. 24–29.
7. Кодирование информации: методические указания / сост. В.Д. Горбоконенко, В.Е. Шикина. – Ульяновск: УлГТУ, 2006. – 56 с.
8. Меняев, М.Ф. Информатика и основы программирования: учеб. пособие для вузов / М.Ф. Меняев. – Москва: Омега-Л, 2007. – ISBN 5-365-00151-6.

9. Мельников, В.П. Информационные технологии: учеб. для вузов / В.П. Мельников. – 2-е изд., стер. – Москва: Академия, 2009. – ISBN978-5-7695-4884-0.
10. Поднебесова, Г.Б. Теоретические основы информатики. Практикум / Г.Б. Поднебесова. – Челябинск: Изд-во Южно-Урал. гос. гуман.-пед. ун-та, 2015. – 92 с. – ISBN 978-5-906777-56-0.
11. Поднебесова, Г.Б. Теоретические основы информатики: учебное пособие / Г.Б. Поднебесова. – Челябинск: Изд-во Южно-Урал. гос. гуман.-пед. ун-та, 2022. – 196 с. – ISBN 978-5-907611-09-2.
12. Подчукаев, В.А. Теория информационных процессов и систем: учеб. пособие для вузов / В.А. Подчукаев. – Москва: Гардарики, 2007. – ISBN 5-8297-0297-5.
13. Распознавание образов и машинное восприятие / А.С. Потапов. – Санкт-Петербург: Политехника, 2007. – ISBN 5-7325-0881-3.
14. Тропченко, А.Ю. Методы сжатия изображений, аудиосигналов и видео / А.Ю. Тропченко, А.А. Тропченко // Теоретическая информатика. – Санкт-Петербург: СПбГУ ИТМО, 2009. – 108 с.
15. Информационно-коммуникационные технологии в образовании. – URL: http://www.ict.edu.ru/lib/ (дата обращения: 16.01.2022).
16. Лидовский, В.В. Теория информации: учеб. пособие / В.В. Лидовский. – Москва: Компания Спутник+, 2004. – 111 с. – ISBN 5-93406-661-7.
17. Теоретическая информатика. – URL: http://se.hse.ru/11003945/teor_inf (дата обращения: 09.04.2022).

ПРИЛОЖЕНИЯ

Приложение 1

Рабочая (модульная) программа

Лекции

Наименование раздела дисциплины (модуля)/ Тема и содержание	***Трудоемкость (кол-во часов)***
1. Информация	***10***
Формируемые компетенции, образовательные результаты: ОПК-1: З.2 (ОПК.1.1), У.1 (ОПК.1.2)	
1.1. Понятие об информации 1. Понятие об информации 2. Различные определения информации 3. Категории свойств информации 4. Формы представления информации.	2
1.2. Данные и знания 1. Данные 2. Свойства декларативных знаний 3. Знания 4. Классификация знаний 5. Модели представления знаний	2
1.3. Каналы передачи данных 1. Канал связи 2. Характеристики каналов связи 3. Условия оптимального использования каналов связи 4. Теорема В.А. Котельникова	2

1.4. Энтропия. Энтропия сложных событий 1. Канал связи 2. Характеристики каналов связи 3. Условия оптимального использования каналов связи	2
1.5. Моделирование данных 1. Основные понятия 2. Метод Баркера 3. Подход, используемый в CASE-средстве SILVERRUN 4. Сетевая теория информации	2
2. Количество информации	*8*
Формируемые компетенции, образовательные результаты: ОПК-1: З.1 (ОПК.1.1), У.1 (ОПК.1.2), У.2 (ОПК.1.2), В.1 (ОПК.1.3) УК-6: У.3 (УК.6.2)	
2.1. Подходы к измерению количества информации 1. Метод Хартли 2. Статистический подход 3. Энтропия 4. Семантический подход 5. Прагматический подход	2
2.2. Количество информации. Общие представления об избыточности 1. Алфавиты 2. Кодирование 3. Системы счисления 4. Количество информации 5. Избыточность. Виды избыточности	2

2.3. Передача дискретных сообщений по каналу без шумов и с шумами 1. Передача сообщений по каналам без шумов 2. Первая теорема Шеннона 3. Обратная теорема Шеннона 4. Передача сообщений по каналам с шумами 5. Вторая теорема Шеннона 6. Помехи и борьба с ними	2
2.4. Информационные системы 1. Понятие информационной системы 2. Образовательные ИС 3. Геоинформационные системы 4. Интеллектуальные информационные системы 5. Экспертные системы в образовании	2
3. Кодирование	***6***
Формируемые компетенции, образовательные результаты: УК-6: З.3 (УК.6.1), У.3 (УК.6.2), В.2 (УК.6.3) ОПК-1: У.2 (ОПК.1.2)	
3.1. Кодирование информации 1. Основные принципы 2. Алфавитное кодирование 3. Минимизация длины кода сообщения 4. Неравенство Макмиллана	2
3.2. Сжатие информации 1. Основные принципы сжатия информации 2. Сжатие с потерями и без потерь 3. Арифметический и вероятностный методы 4. Метод повторяющихся последовательностей RLE 5. Метод словарей. Метод Лемпеля–Зива, LZ78	2

3.3. Криптография. Электронная подпись 1. Методы защиты данных 2. Электронная цифровая подпись 3. Хэш-функции 4. Криптосистема Эль-Гамаля	2
4. Помехоустойчивое кодирование	***6***
Формируемые компетенции, образовательные результаты: УК-6: З.3 (УК.6.1) ОПК-1: У.2 (ОПК.1.2)	
4.1. Помехоустойчивое кодирование 1. Принципы помехоустойчивого кодирования 2. Помехи 3. Классификация помехоустойчивых кодов 4. Кодовое расстояние 5. Код Хемминга	2
4.2. Помехоустойчивое кодирование в системах сотовой связи 1. Виды помех и борьба с ними в системах сотовой связи 2. Интерливинг. Перемежение 3. Адаптивная коррекция 4. Стандарты сотовой связи	2
4.3. Мировые информационные ресурсы и глобальные информационные сети 1. Информационные ресурсы 2. Классификация мировых информационных ресурсов 3. Государственные информационные ресурсы	

Наименование раздела дисциплины (модуля)/ Тема и содержание	*Трудоемкость (кол-во часов)*
1. Информация	***12***
Формируемые компетенции, образовательные результаты: ОПК-1: З.2 (ОПК.1.1), У.1 (ОПК.1.2)	
1.1. Вычисление статистических характеристик текстовой информации 1. Определение количества информации 2. Построение таблицы частот 3. Анализ частоты появления букв русского алфавита в тексте с помощью Excel	2
1.2. Работа с псевдослучайными числами и оценка их качества статистическими тестами 1. Применение функций Excel для получения случайных чисел 2. Работа с программой Псевдослучайные последовательности 3. Рассмотрение систем оценки качества генераторов псевдослучайных последовательностей	2
1.3. Разработка программы формирования псевдослучайных чисел 1. Разработка программы на C# для оценки качества полученных последовательностей 2. Осуществить проверку на равномерность распределения	4

1.4. Моделирование данных 1. Подготовить сообщение по одной из тем 2. Создать интеллект карту	4
2. Количество информации	***6***
Формируемые компетенции, образовательные результаты: ОПК-1: З.1 (ОПК.1.1), У.1 (ОПК.1.2), У.2 (ОПК.1.2), В.1 (ОПК.1.3) УК-6: У.3 (УК.6.2)	
2.1. Разработка программы подсчета количества информации различными методами 1. Методы К. Шеннона и Р. Хартли 2. Реализация алгоритма вычисления количества информации 3. Метод А.А. Харкевича 4. Реализация алгоритма вычисления количества прагматической информации	6
3. Кодирование	***10***
Формируемые компетенции, образовательные результаты: УК-6: З.3 (УК.6.1), У.3 (УК.6.2), В.2 (УК.6.3) ОПК-1: У.2 (ОПК.1.2)	
3.1. Кодирование и декодирование символьной информации с использованием различных кодовых таблиц 1. Различные кодовые таблицы 2. Кодовые таблицы ASCII 3. Кодирование символьной информации	2

3.2. Кодирование графической, звуковой и видеоинформации 1. Двоичное кодирование графической информации 2. Дискретизация и квантование 3. Кодирование видеоинформации 4. Метод JPEG. Алгоритм MPEG 5. Примеры кодирования звуковой и видео информации	2
3.3. Кодирование информации методом Шеннона–Фано 1. Кодирование информации 2. Методом Шеннона–Фано закодировать фразу «Сшит колпак да не поколпаковски»	2
3.4. Кодирование информации методом Хаффмана 1. Кодирование информации 2. Методом Хаффмана закодировать фразу «На дворе трава, на траве дрова»	2
3.5. Сжатие данных по методу Лемпеля–Зива 1. Закодировать фразу «прямпрямпрямопрямо» LZ-методом 2. Применить алгоритм LZ78 для кодирования последовательности 0100011101010	2
4. Помехоустойчивое кодирование	***6***
Формируемые компетенции, образовательные результаты: УК-6: З.3 (УК.6.1) ОПК-1: У.2 (ОПК.1.2)	

4.1. Коды Хемминга, исправляющие одиночную ошибку 1. Обнаружение одиночной ошибки (1-й метод) – применить три проверки 2. Выполнение сложения по модулю 2 3. Обнаружение одиночной ошибки (2-й метод)	2
4.2. Алгоритмы цифровой подписи 1. Шифр Эль-Гамаля. Описание метода 2. Алгоритм цифровой подписи DSA. Пример	2
4.3. Защита кейса (помехоустойчивые коды) 1. Разработать кейс 2. Описать предложенный метод кодирования, привести 2 примера	2

Приложение 2

Содержание самостоятельной работы

Наименование раздела дисциплины (модуля)/ Тема для самостоятельного изучения	*Трудоемкость (кол-во часов)*
1. Информация	***20***
Формируемые компетенции, образовательные результаты: ОПК-1: З.2 (ОПК.1.1), У.1 (ОПК.1.2)	
1.1. Понятие об информации ***Задание для самостоятельного выполнения студентом:*** Подготовить сообщение. Показать вклад конкретного ученого в развитие теории информации (номер по списку)	10
1.2. Моделирование данных ***Задание для самостоятельного выполнения студентом:*** Подготовить презентацию по одной из тем: 1. Кодирование сети 2. Канал с множественным доступом 3. Широковещательный канал 4. Relay Channel 5. Interference channel 6. Когнитивное радио 7. Масштабирование сети 8. Портфельная теория 9. Универсальное кодирование источника	10

2. Количество информации	***20***
Формируемые компетенции, образовательные результаты: ОПК-1: З.1 (ОПК.1.1), У.1 (ОПК.1.2), У.2 (ОПК.1.2), В.1 (ОПК.1.3) УК-6: У.3 (УК.6.2)	
2.1. Информационные системы ***Задание для самостоятельного выполнения студентом:*** 1. Системы поддержки принятия решений 2. Логистические ИС 3. ИС «Сетевой город» 4. ИС, ускоряющие поток товаров 5. Системы автоматизированного проектирования 6. Информационно-справочные системы 7. Медицинские ИС 8. Информационные системы образования 9. ИС по отысканию рыночной ниши 10. ИС GEO 11. Информационно-вычислительные системы 12. Государственные ИС	20
3. Кодирование	***10***
Формируемые компетенции, образовательные результаты: УК-6: З.3 (УК.6.1), У.3 (УК.6.2), В.2 (УК.6.3) ОПК-1: У.2 (ОПК.1.2)	
3.1. Кодирование информации ***Задание для самостоятельного выполнения студентом:*** 1. Закодировать фразу «Сшит колпак да не поколпаковски, надо колпак переколпаковать». Использовать метод Шеннона–Фано	10

2. Закодировать фразу «всем всем всем и каждому скажу», используя код LZ78	
4. Помехоустойчивое кодирование	30
Формируемые компетенции, образовательные результаты: УК-6: З.3 (УК.6.1) ОПК-1: У.2 (ОПК.1.2)	
Защита кейса (помехоустойчивые коды) ***Задание для самостоятельного выполнения студентом:*** 1. Передан код числа 7 в виде «0 1 1 0 1 0 0», а приняли в виде «1 1 1 0 1 0 0». Обнаружить позицию ошибки 2. Основной код 0101100, принятый код 0101101. Подсчитать дополнительный код. Обнаружить позицию ошибки	30

Приложение 3

Рейтинг

Фамилия Имя	Тест1		Доп.	Анализ текстов		Статистич. тесты лб		Тестирование ГСЧ		Модуль1	Тест2		Доп.	Моделирование данных		Изм . кол. инф.		Метод Харкевича+		Модуль2
	"Вес"	Коэф.		"Вес"	Коэф	Вес	Коэф.	Вес	Коэф.		"Вес"	Коэф.		"Вес"	Коэф.	"Вес"	Коэф.	Вес	Коэф.	
Асабин Владимир	50	0,69	0,05	15	1	15	1	20	1	87	50	0,75	0,05	10	1	20	1	20	1	90

Тест3		Доп.	Кодирование информации		Кодирование ГЗВ инф.		Шеннона-Фено		Хаффмена		Лемпеля-Зива		Модуль3	Тест4		Доп.	Помехоуст. кодир-е		Хемминг лаб.		ЭЦП лаб.		Модуль4	Итого	Оценка
"Вес"	Коэф.		"Вес"	Коэф.	"Вес"	Коэф.	"Вес"	Коэф	"Вес"	Коэф	"Вес"	Коэф		"Вес"	Коэф.		"Вес"	Коэф.	"Вес"	Коэф	"Вес"	Коэф			
50	0,74	0,09	5	1	5	1	10	1	15	1	15	1	91,5	60	0,96	0,03	14	1,2	13	1,2	13	1,2	107,4	93,35	отлично

Приложение 4

Анализ текстов

Анализ текстовой информации в настоящее время является актуальной задачей. Применение статистики для анализа текстов – традиционная задача[1].

Рассмотрим некоторые интересные факты относительно частоты встречаемости букв и их сочетаний в разных языках (подробнее см. например, недавно вышедшую интересную книгу Анализ текстов[2]).

Частотные характеристики текстовых сообщений

Текст состоит из слов, слова из букв. Количество различных букв в каждом языке ограничено и буквы могут быть просто перечислены. Существуют такие характеристики текста, как повторяемость букв, пар букв (биграмм) и вообще m-ок (m-грамм), а также, сочетаемость букв друг с другом, чередование гласных и согласных и некоторые другие. Замечательно, что эти характеристики являются достаточно устойчивыми.

Используя систему *STATISTICA* можно проверить эти закономерности, например, в текстах Интернет. Идея состоит в подсчете чисел вхождений каждой n^m возможных m-грамм в достаточно длинных открытых текстах $T = t_1 t_2 \ldots t_l$, составленных из букв алфавита $\{a_1, a_2, \ldots, a_n\}$. При этом просматриваются подряд идущие m-граммы текста:

$$t_1 t_2 \ldots t_m,\ t_2 t_3 \ldots t_{m+1},\ \ldots,\ t_{i-m+1} t_{l-m+2} \ldots t_l.$$

---

[1] Анализ текстов. – URL: http://statistica.ru/local-portals/data-mining/analiz-tekstov/.

[2] Алферов, А.П. Криптография / А.П. Алферов, А.Ю. Зубов, А.С. Кузьмин, А.В. Черемушкин. – Москва: Гелиос АРВ, 2002.

Если $\theta(a_{i1}, a_{i2}, \dots, a_{im})$ – число появлений m-граммы $a_{i1}a_{i2}...a_{im}$ в тексте T, а L – общее число подсчитанных m-грамм, то опыт показывает, что при достаточно больших L частоты

$$\frac{\theta(a_{i1}, a_{i2}, \dots, a_{im})}{L}$$

для данной m-граммы мало отличаются друг от друга.

В силу этого, относительную частоту считают приближением вероятности $P(a_{i1}a_{i2}...a_{im})$ появления данной m-граммы в случайно выбранном месте текста (такой подход принят при статистическом определении вероятности).

Ниже приводится таблицы частот букв (в процентах) ряда европейских языков. Данные заимствованы из книги Elements de cryptographie[3].

Некоторая разница значений частот в приводимых в различных источниках таблицах объясняется тем, что частоты существенно зависят не только от длины текста, но и от его характера. Например, в технических текстах редкая буква *Ф* может стать довольно частой в связи с частым использованием таких слов, как функция, дифференциал, диффузия, коэффициент и т.п.

Еще большие отклонения от нормы в частоте употребления отдельных букв наблюдаются в некоторых художественных произведениях, особенно в стихах. Поэтому для надежного определения средней частоты букв желательно иметь набор различных текстов, заимствованных из различных источников. Вместе с тем, как правило, подобные отклонения незначительны, и в первом приближении ими можно пренебречь.

---

[3] Baudouin, C. Elements de cryptographie / C. Baudouin. – Paris, 1939.

Таблица 1 – Таблицы частот букв европейских языков

Буква алфавита	Французский язык	Немецкий язык	Английский язык	Испанский язык	Итальянский язык
A	7.68	5.52	7.96	12.90	11.12
B	0.80	1.56	1.60	1.03	1.07
C	3.32	2.94	2.84	4.42	4.11
D	3.60	4.91	4.01	4.67	3.54
E	17.76	19.18	12.86	14.15	11.63
F	1.06	1.96	2.62	0.70	1.15
G	1.10	3.60	1.99	1.00	1.73
H	0.64	5.02	5.39	0.91	0.83
I	7.23	8.21	7.77	7.01	12.04
J	0.19	0.16	0.16	0.24	–
K	–	1.33	0.41	–	–
L	5.89	3.48	3.51	5.52	5.95
M	2.72	1.69	2.43	2.55	2.65
N	7.61	10.20	7.51	6.20	7.68
O	5.34	2.14	6.62	8.84	8.92
P	3.24	0.54	1.81	3.26	2.66
Q	1.34	0.01	0.17	1.55	0.48
R	6.81	7.01	6.83	6.95	6.56
S	8.23	7.07	6.62	7.64	4.81
T	7.30	5.86	9.72	4.36	7.07
U	6.05	4.22	2.48	4.00	3.09
V	1.27	0.84	1.15	0.67	1.67
W	–	1.38	1.80	–	–
X	0.54	–	0.17	0.07	–
Y	0.21	–	1.52	1.05	–
Z	0.07	1.17	0.05	0.31	1.24

Наглядное представление о частотах букв дает диаграмма встречаемости. Так, для английского языка, в соответствии с таблицей, такая диаграмма изображена на рис. 1. Для ее построения использована система *STATISTICA*.

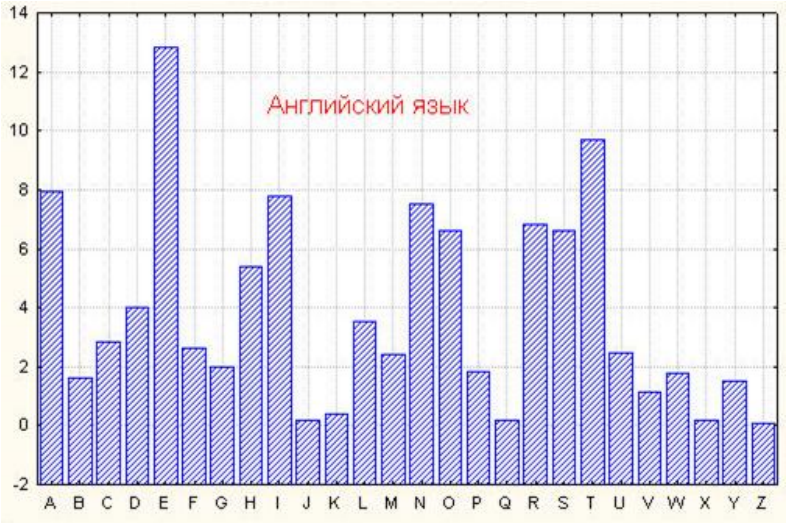


Рис. 1. Диаграмма частоты использования букв алфавита в английском языке

Для русского языка частоты (в порядке убывания) знаков алфавита, в котором отождествлены *Е* с *Ё*, *Ь* с *Ъ*, а также имеется знак пробела (–) между словами, приведены в следующей таблице (см. в книге Вероятность и информация[4]).

Таблица 2 – Частота знаков алфавита для русского языка

- 0.175	**О** 0.090	**Е, Ё** 0.072	**А** 0.062
И 0.062	**Т** 0.053	**Н** 0.053	**С** 0.045
Р 0.040	**В** 0.038	**Л** 0.035	**К** 0.028
М 0.026	**Д** 0.025	**П** 0.023	**У** 0.021

[4] Яглом, А.М. Вероятность и информация / А.М. Яглом, И.М. Яглом. – Москва: Наука, 1973.

Окончание таблицы 2

Я 0.018	**Ы** 0.016	**З** 0.016	**Ь, Ъ** 0.014
Б 0.014	**Г** 0.013	**Ч** 0.012	**Й** 0.010
Х 0.009	**Ж** 0.007	**Ю** 0.006	**Ш** 0.006
Ц 0.004	**Щ** 0.003	**Э** 0.003	**Ф** 0.002

На основании таблицы получаем следующую диаграмму частот (рис. 2).

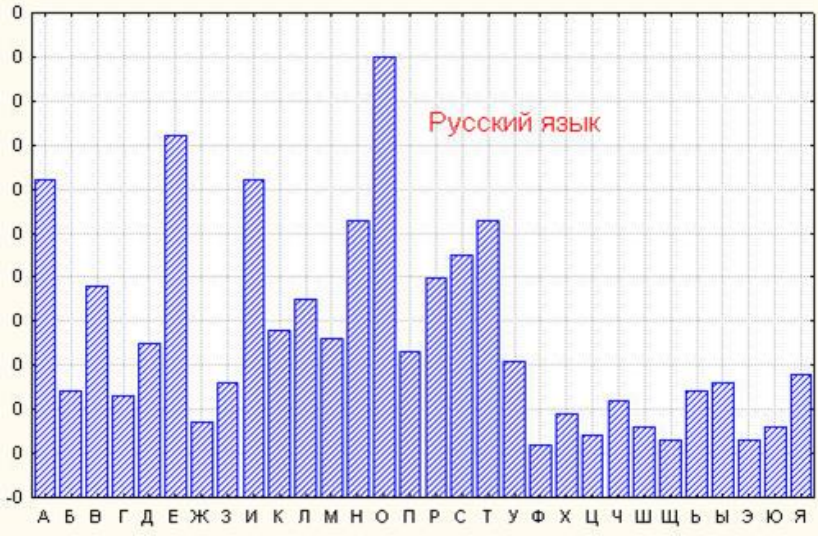


Рис. 2. Диаграмма частот использования букв алфавита для русского языка

Имеется мнемоническое правило запоминания десяти наиболее частых букв русского алфавита. Эти буквы составляют нелепое слово СЕНОВАЛИТР. Можно также

предложить аналогичный способ запоминания частых букв английского языка, например, с помощью слова TETRIS-HONDA (см. таблицу).

Таблица 3 – Правило запоминания десяти наиболее часто используемых букв

Французский язык	E, S, A, N, T, I, R, U, L, O	79.9%
Немецкий язык	E, N, I, S, T, A, H, D, U	77.2%
Английский язык	E, T, A, I, N, R, O, S, H, D	75.3%
Испанский язык	E, A, O, S, I, R, N, L, D, C	78.3%
Итальянский язык	I, E, A, O, N, T, R, L, S, T	79.9%

Устойчивыми являются также частотные характеристики биграмм, триграмм и четырехграмм осмысленных текстов.

Приведем таблицы частот биграмм для русского языка (таблицы заимствованы из книги Military cryptanalysis[5]). Для удобства они разбиты на четыре части (см. таблицу 4).

[5] Friedman, W.F. Military cryptanalysis / W.F. Friedman, D. Callimahos. – Part 1, Vol 2. – Aegean Park Press. Laguna Hills CA, 1985.

Таблица 4 – Таблицы частот биграмм для русского языка

Часть 1																
	А	**Б**	**В**	**Г**	**Д**	**Е**	**Ж**	**З**	**И**	**Й**	**К**	**Л**	**М**	**Н**	**О**	**П**
А	2	12	35	8	14	7	6	15	7	7	19	27	19	45	5	11
Б	5					9	1		6			6		2	21	
В	35	1	5	3	3	32		2	17		7	10	3	9	58	6
Г	7				3	3			5		1	5		1	50	
Д	25		3	1	1	29	1	1	13		1	5	1	13	22	3
Е	2	9	18	11	27	7	5	10	6	15	13	35	24	63	7	16
Ж	5	1			6	12			5					6		
З	35	1	7	1	5	3			4		2	1	2	9	9	1
И	4	6	22	5	10	21	2	23	19	11	19	21	20	32	8	13
Й	1	1	4	1	3		1	2	4		5	1	2	7	9	7
К	24	1	4	1		4	1	1	26		1	4	1	2	66	2
Л	25	1	1	1	1	33	2	1	36		1	2	1	8	30	2
М	18	2	4	1	1	21	1	2	23		3	1	3	7	19	5
Н	54	1	2	3	3	34			58		3		1	24	67	2
О	1	28	84	32	47	15	7	18	12	29	19	41	38	30	9	18
П	7					15			4			9		1	46	

Часть 2															
	Р	**С**	**Т**	**У**	**Ф**	**Х**	**Ц**	**Ч**	**Ш**	**Щ**	**Ы**	**Ь**	**Э**	**Ю**	**Я**
А	26	31	27	3	1	10	6	7	10	1			2	6	9
Б	8	1		6						1	11				2
В	6	19	6	7		1	1	2	4	1	18	1	2		3
Г	7			2											
Д	6	8	1	10			1	1	1		5	1			1
Е	39	37	33	3	1	8	3	7	3	3			1	1	2
Ж		1													
З	3	1		2							4				4
И	11	29	29	3	1	17	3	11	1	1			1	3	17
Й	3	10	2				1	3	2						
К	10	3	7	10			1								
Л		3	1	6		4		1			3	20		4	9
М	2	5	3	9	1			2			5	1	1		3
Н	1	9	9	7	1		5	2			36	3			5
О	43	50	39	3	2	5	2	12	4	3			2	3	2
П	41	1		6							2				2

Часть 3																
	А	**Б**	**В**	**Г**	**Д**	**Е**	**Ж**	**З**	**И**	**Й**	**К**	**Л**	**М**	**Н**	**О**	**П**
Р	55	1	4	4	3	37	3	1	24		3	1	3	7	56	2
С	8	1	7	1	2	25			6		40	13	3	9	27	11
Т	35	1	27	1	3	31		1	28		5	1	1	11	56	4
У	1	4	4	4	11	2	6	3	2		8	5	5	5	1	5
Ф	2					2			2						1	
Х	4	1	4	1	3	1		2	3		4	3	3	4	18	5
Ц	3					7			10		2				1	
Ч	12					23			13		2			6		
Ш	5					11			14		1	2		2	2	
Щ	3					8			6					1		
Ы		1	9	1	3	12		2	4	7	3	6	6	3	2	10
Ь		2	4	1	1	2		2	2		6		3	13	2	4
Э											1			1		
Ю		2	1	2	1			3	1		1		1	1	1	3
Я	1	3	9	1	3	3	1	5	3	2	3	3	4	6	3	6

Часть 4															
	Р	**С**	**Т**	**У**	**Ф**	**Х**	**Ц**	**Ч**	**Ш**	**Щ**	**Ы**	**Ь**	**Э**	**Ю**	**Я**
Р	1	5	9	16		1	1	1	2		8	3			5
С	4	11	82	6		1	1	2	2		1	8			17
Т	26	18	2	10				1			11	21			4
У	7	14	7			1		8	3	2				9	1
Ф	1	1													
Х	3	4	2	2	1			1							
Ц				1							1				
Ч			7	1					1			1			
Ш				1								1			
Щ				1											
Ы	3	9	4	1		16		1	2						
Ь	1	11	3					1	4				1	3	1
Э		1	9												
Ю	1	1	7				1	1		4					
Я	3	6	10			2	1	4	1	1			1	1	1

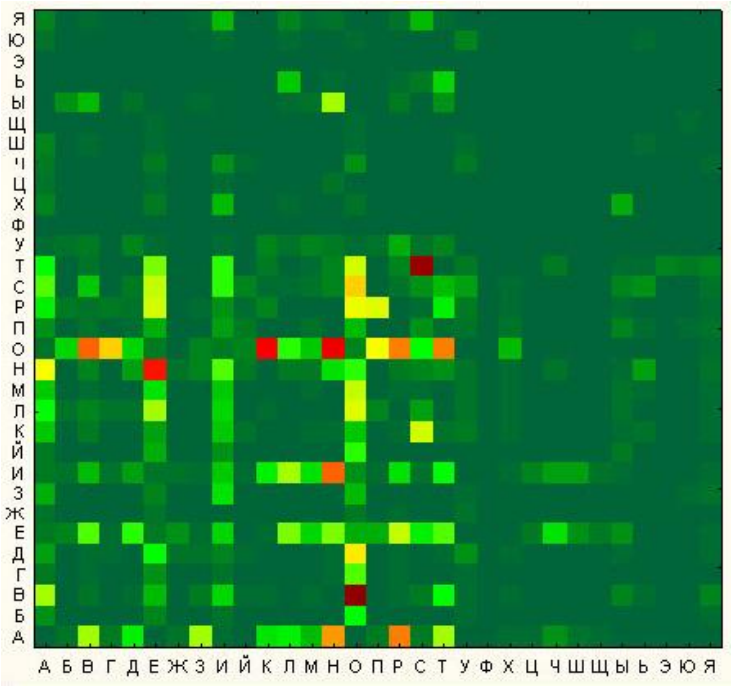


Рис. 3. График частот использования пар букв русского алфавита

Для получения более точных сведений об открытых текстах можно строить и анализировать таблицы *k*-грамм при $k > 2$, однако для учебных целей вполне достаточно ограничиться биграммами. Неравномерность *k*-грамм (и даже слов) тесно связана с характерной особенностью открытого текста – наличием в нем большого числа повторений отдельных фрагментов текста: корней, окончаний, суффиксов, слов и фраз. Так, для русского языка такими привычными фрагментами являются наиболее частые биграммы и триграммы:

СТ, НО, ЕН, ТО, НА, ОВ, НИ, РА, ВО, КО
СТО, ЕНО, НОВ, ТОВ, ОВО, ОВА

Полезной является информация о сочетаемости букв, то есть о предпочтительных связях букв друг с другом, которую легко извлечь из таблиц частот биграмм.

Имеется в виду таблица, в которой слева и справа от каждой буквы расположены наиболее предпочтительные «соседи» (в порядке убывания частоты соответствующих биграмм). В таких таблицах обычно указывается также доля гласных и согласных букв (в процентах), предшествующих (или следующих за) данной букве.

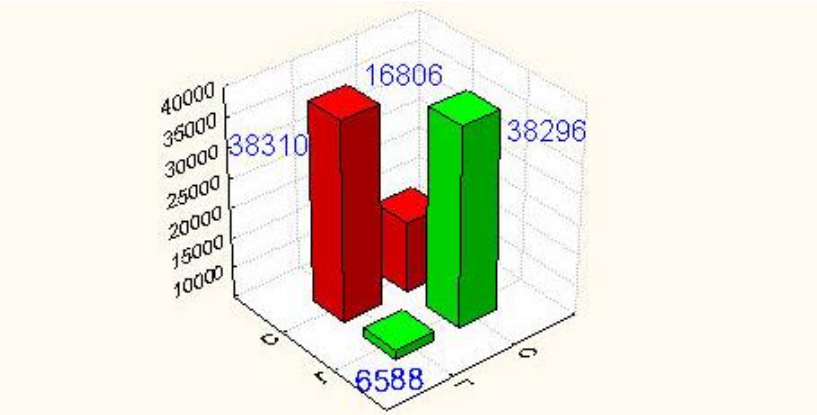


Рис. 4. Частота чередования гласных и согласных букв

При анализе сочетаемости букв друг с другом следует иметь в виду зависимость появления букв в открытом тексте от значительного числа предшествующих букв. Для анализа этих закономерностей используют понятие условной вероятности.

Наблюдения над открытыми текстами показывают, что для условных вероятностей выполняются неравенства $p(a_{i1}) \;\&\; ne;\; p(a_{i1}/a_{i2}),\; p(a_{i1}/a_{i2}) \;\&\; ne;\; p(a_{i1}/a_{i2}a_{i3}),$ ....

Таблица 5 – Сочетаемость букв русского языка

Г	С	Слева		Справа	Г	С
3	97	л, д, к, т, в, р, н	**А**	л, н, с, т, р, в, к, м	12	88
80	20	я, е, у, и, а, о	**Б**	о, ы, е, а, р, у	81	19
68	32	я, т, а, е, и, о	**В**	о, а, и, ы, с, н, л, р	60	40
78	22	р, у, а, и, е, о	**Г**	о, а, р, л, и, в	69	31
72	28	р, я, у, а, и, е, о	**Д**	е, а, и, о, н, у, р, в	68	32
19	81	м, и, л, д, т, р, н	**Е**	н, т, р, с, л, в, м, и	12	88
83	17	р, е, и, а, у, о	**Ж**	е, и, д, а, н	71	29
89	11	о, е, а, и	**З**	а, н, в, о, м, д	51	49
27	73	р, т, м, и, о, л, н	**И**	с, н, в, и, е, м, к, з	25	75
55	45	ь, в, е, о, а, и, с	**К**	о, а, и, р, у, т, л, е	73	27
77	23	г, в, ы, и, е, о, а	**Л**	и, е, о, а, ь, я, ю, у	75	25
80	20	я, ы, а, и, е, о	**М**	и, е, о, у, а, н, п, ы	73	27
55	45	д, ь, н, о	**Н**	о, а, и, е, ы, н, у	80	20
11	89	р, п, к, в, т, н	**О**	в, с, т, р, и, д, н, м	15	85
65	35	в, с, у, а, и, е, о	**П**	о, р, е, а, у, и, л	68	32
55	45	и, к, т, а, п, о, е	**Р**	а, е, о, и, у, я, ы, н	80	20
69	31	с, т, в, а, е, и, о	**С**	т, к, о, я, е, ь, с, н	32	68
57	43	ч, у, и, а, е, о, с	**Т**	о, а, е, и, ь, в, р, с	63	37
15	85	п, т, к, д, н, м, р	**У**	т, п, с, д, н, ю, ж	16	84
70	30	н, а, е, о, и	**Ф**	и, е, о, а, е, о, а	81	19
90	10	у, е, о, а, ы, и	**Х**	о, и, с, н, в, п, р	43	57
69	31	е, ю, н, а, и	**Ц**	и, е, а, ы	93	7
82	18	е, а, у, и, о	**Ч**	е, и, т, н	66	34
67	33	ь, у, ы, е, о, а, и, в	**Ш**	е, и, н, а, о, л	68	32
84	16	е, б, а, я, ю	**Щ**	е, и, а	97	3
0	100	м, р, т, с, б, в, н	**Ы**	л, х, е, м, и, в, с, н	56	44
0	100	н, с, т, л	**Ь**	н, к, в, п, с, е, о, и	24	76
14	86	с, ы, м, л, д, т,, р, н	**Э**	н, т, р, с, к	0	100
58	42	ь, о, а, и, л, у	**Ю**	д, т, щ, ц, н, п	11	89
43	57	о, н, р, л, а, и, с	**Я**	в, с, т, п, д, к, м, л	16	84

Систематически вопрос о зависимости букв алфавита в открытом тексте от предыдущих букв исследовался известным русским математиком А.А. Марковым (1856–1922). Он доказал, что появления букв в открытом тексте нельзя считать независимыми друг от друга. В связи с этим А.А. Марковым отмечена еще одна устойчивая закономерность открытых текстов, связанная с чередованием гласных и согласных букв. Им были подсчитаны частоты встречаемости биграмм вида гласная-гласная (*г*, *г*), гласная-согласная (*г*, *с*), согласная-гласная (*с*, *г*), согласная-согласная (*с*, *с*) в русском тексте длиной в 10^5 знаков. Результаты подсчета отражены в следующей таблице 6.

Таблица 6 – Частота встречаемости биграмм вида гласная–гласная

	Г	**С**	**Всего**
Г	6588	38310	44898
С	38296	16806	55102

Из этой таблицы видно, что для русского языка характерно чередование гласных и согласных, причем относительные частоты могут служить приближениями соответствующих условных и безусловных вероятностей:

$$p(г/с) \approx 0.663,\ p(с/г) \approx 0.872,\ p(г) \approx 0.432,\ p(с) \approx 0.568.$$

После А.А. Маркова зависимость появления букв текста вслед за несколькими предыдущими исследовал методами теории информации К. Шеннон. Фактически им было показано, в частности, что такая зависимость ощутима на глубину приблизительно в 30 знаков, после чего она практически отсутствует.

Таблица 7 – Доля гласных букв в литературном тексте

Французский язык	44.27%
Немецкий язык	39.27%
Английский язык	39.21%
Испанский язык	47.95%
Итальянский язык	46.80%

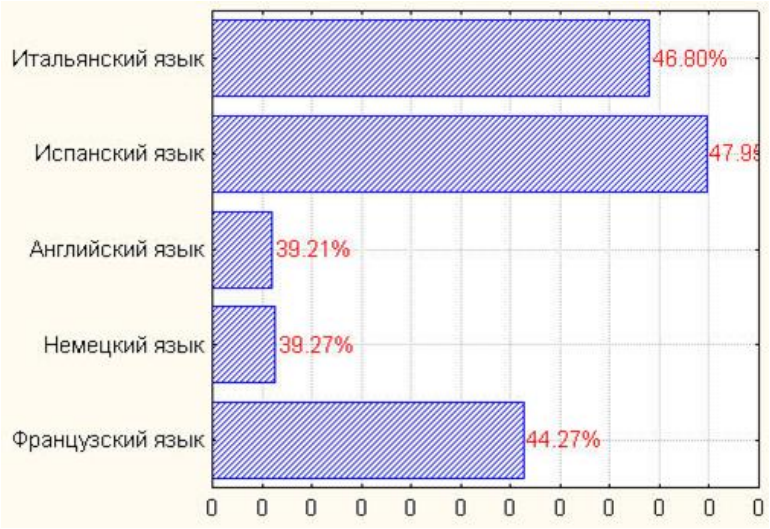


Рис. 4. Доля гласных букв в алфавитах

Приведенные выше закономерности имеют место для обычных «читаемых» открытых текстов, используемых при общении людей. Как уже отмечалось ранее, эти закономерности играют большую роль в криптоанализе. В частности, они используются при построении формализованных критериев на открытый текст, позволяющих применять методы математической статистики в задаче распознавания открытого текста в потоке сообщений. При использовании

же специальных алфавитов требуются аналогичные исследования частотных характеристик «открытых текстов», возникающих, например, при межмашинном обмене информацией или в системах передачи данных. В этих случаях построение формализованных критериев на «открытый текст» – задача значительно более сложная.

Помимо криптографии частотные характеристики открытых сообщений существенно используются и в других сферах. Например, клавиатура компьютера, пишущей машинки или линотипа – это замечательное воплощение идеи ускорения набора текста, связанное с оптимизацией расположения букв алфавита относительно друг друга в зависимости от частоты их применения.

Приложение 5

Пример текста для анализа

Information Theory[6]

Imagine that someone hands you a sealed envelope, containing, say, a telegram. You want to know what the message is, but you can't just open it up and read it. Instead, you have to play a game with the messenger: you get to ask yes-or-no questions about the contents of the envelope, to which he'll respond truthfully. Question: assuming this rather contrived and boring exercise is repeated many times over, and you get as clever at choosing your questions as possible, what's the smallest number of questions needed, on average, to get the contents of the message nailed down?

This question actually has an answer. Suppose there are only a finite number of messages («Yes»; «No»; «Marry me?»; «In Reno, divorce final»; «All is known stop fly at once stop»; or just that there's a limit on the length of the messages, say a thousand characters). Then we can number the messages from 1 to N. Call the message we get on this trial *S*. Since the game is repeated many times, it makes sense to say that there's a probability p_i of getting message number *i* on any given trial, i.e. Prob (*S* = i) = p_i. Now, the number of yes-no questions needed to pick out any given message is, at most, *log N*, taking the logarithm to base two. (If you were allowed to ask questions with three possible answers, it'd be log to the base three. Natural logarithms would seem to imply the idea of their being 2.718... Answers per question, but nonetheless make sense

[6] Information Theory. – URL: http://bactra.org/notebooks/information-theory.html.

mathematically). But one can do better than that: if message *i* is more frequent than message *j* (if $p_i > p_j$), it makes sense to ask whether the message is *i* before considering the possibility that it's *j*; you'll save time. One can in fact show, with a bit of algebra, that the smallest average number of yes-no questions is: $-\sum_{i=1}^{n} p_i \log p_i$.

This gives us *log N* when all the p_i are equal, which makes sense: then there are no prefered messages, and the order of asking doesn't make any difference. The sum is called, variously, the information, the information content, the self-information, the entropy or the Shannon entropy of the message, conventionally written *H*[*S*].

Now, at this point a natural and sound reaction would be to say «the mathematicians can call it what they like, but what you've described, this ridiculous guessing game, has squat-all to do with information». Alas, would that this were so: it *is* ridiculous, but it works. More: it was arrived at, simultaneously, by several mathematicians and engineers during World War II (among the Americans, most notably, Claude Shannon and Norbert Wiener), working on very serious and practical problems of coding, code-breaking, communication and automatic control. The real justification for regarding the entropy as the amount of information is that, unsightly though it is, though it's abstracted away all the content of the message and almost all of the context (except for the distribution over messages), it works. You can try to design a communication channel which doesn't respect the theorems of information theory; in fact, people did; you'll fail, as they did.

Учебное издание

Поднебесова Галина Борисовна

ТЕОРИЯ ИНФОРМАЦИИ. ДАННЫЕ. ЗНАНИЯ
Учебно-практическое пособие

Работа рекомендована РИС ЮУрГГПУ
Протокол № 26 от 2022 г.

ISBN 978-5-907611-46-7

Издательство ЮУрГГПУ
454080, г. Челябинск, пр. Ленина, 69

Редактор Е.М. Сапегина
Технический редактор А.Г. Петрова

Объем 2,4 уч.-изд. л. (6,45 усл. печ. л.)
Подписано в печать 24.08.2022 г.
Тираж 100 экз. Бумага типографская
Формат 60x84/16 Заказ №

Отпечатано с готового оригинал-макета
в типографии ЮУрГГПУ
454080, г. Челябинск, пр. Ленина, 69