

Г.Б. Поднебесова

**ТЕОРЕТИЧЕСКИЕ
ОСНОВЫ
ИНФОРМАТИКИ**

Практикум

Министерство образования и науки Российской Федерации
Федеральное государственное бюджетное образовательное
учреждение высшего профессионального образования
«ЧЕЛЯБИНСКИЙ ГОСУДАРСТВЕННЫЙ ПЕДАГОГИЧЕСКИЙ УНИВЕРСИТЕТ»

Г.Б. Поднебесова

**ТЕОРЕТИЧЕСКИЕ ОСНОВЫ
ИНФОРМАТИКИ
ПРАКТИКУМ**

Челябинск
2015

УДК 001.8 (076)

ББК 73я7

П 44

Поднебесова, Г.Б. Теоретические основы информатики [Текст]: практикум / Г.Б. Поднебесова. – Челябинск: Изд-во Челяб. гос. пед. ун-та, 2015. – 109 с.

ISBN 978-5-906777-56-0

В практикуме представлены материалы для изучения курсов «Теоретические основы информатики», «Теория информации» и «Теоретические основы информатики и современных информационных технологий». Пособие предназначено для организации аудиторной и самостоятельной работы студентов, обучающихся по направлениям «Педагогическое образование» и «Информационные системы и технологии».

Структура пособия позволяет использовать модульно-рейтинговую систему оценивания учебных достижений студентов при изучении дисциплины «Теоретические основы информатики». В работе имеется также банк тестовых заданий. Издание данного учебного пособия поддержано грантом № 2807 Министерства образования и науки Российской Федерации.

Практикум адресован преподавателям и учителям, для которых интересна данная предметная область.

Рецензенты: В.В. Карачик, д-р физ.-мат. наук, профессор
А.А. Рузаков, канд. пед. наук

ISBN978-5-906777-56-0

© Г.Б. Поднебесова, 2015

© Издательство Челябинского государственного педагогического университета, 2015

Содержание

Введение.....	4
Содержание разделов дисциплины.....	7
Темы и планы лабораторных занятий	9
Модуль 1. Теория информации.....	11
Модуль 2. Кодирование	24
Модуль 3. Автоматы.....	41
Модуль 4. Грамматики.....	
Модуль 5. Распознавание образов. Математическая кибернетика.....	61
Заключение.....	79
Приложения	
Приложение 1 Рабочая (модульная) программа.....	81
Приложение 2. Содержание самостоятельной работы	90
Приложение 3. Рейтинг.....	
Приложение 4. Анализ текстов	92
Приложение 5. Пример текста для анализа.....	107

ВВЕДЕНИЕ

Стратегия развития современного общества на основе знаний и высокоэффективных технологий потребовала внесения значительных коррективов 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.

Учебные дисциплины «Теоретические основы информатики», «Теория информации» базируются на материале предшествующих ей дисциплин математики (Математический анализ, Алгебра и теория чисел), курсов по Абстрактной и компьютерной алгебре и Теории алгоритмов.

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.

Изучение данной дисциплины предусмотрено вариативной частью профессионального цикла основной образовательной программы ФГОС ВПО.

Основной компетенцией, формируемой в результате освоения дисциплины, является способность использовать знания о современной естественнонаучной картине мира в образовательной и профессиональной деятельности, применять методы математической обработки информации, теоретического и экспериментального исследования (ОК-4).

Дополнительная компетенция, формируемая в процессе изучения данного курса, – готовность применять знания теоретической информатики, фундаментальной и прикладной математики для анализа и синтеза информационных систем и процессов (СК-1).

Таблица 1

Планируемые результаты обучения

№ п/п	Компетенция (содержание и обозначение в соответствии с ФГОС ВПО и ООП)	Конкретизированные цели освоения дисциплины		
		знать	уметь	владеть
1	способность использовать знания о современной естественнонаучной картине мира в образовательной и профессиональной деятельности, применять методы математической обработки информации, теоретического и экспериментального исследования (ОК-4)	3.1 иметь представление об общих проблемах и задачах теоретической информатики; 3.2 основные подходы к измерению количества информации; 3.3 способы кодирования информации	У.1 применять алгоритмы для кодирования информации; У.2 применять методы помехоустойчивого кодирования информации	В.1 способами анализа текста; В.2 методами кодирования информации
2	готовность применять знания теоретической информатики, фундаментальной и прикладной математики для анализа и синтеза информационных систем и процессов (СК-1)	3.4 сущность теории автоматов; 3.5 способы построения грамматик; 3.6 основные принципы распознавания образов	У.3 строить регулярные выражения; У.4 применять алгоритмы для распознавания образов	В.3 способами построения автоматов; В.4 способами построения диаграмм Вирта; В.5 методами грамматического разбора

СОДЕРЖАНИЕ РАЗДЕЛОВ ДИСЦИПЛИНЫ

1. ПРЕДМЕТ ТЕОРЕТИЧЕСКОЙ ИНФОРМАТИКИ.

ОСНОВНЫЕ ПОНЯТИЯ

Предмет информатики. Информатика как наука и как вид практической деятельности. Место информатики в системе наук. Роль информации в современном обществе. Виды информационных процессов. Принципы получения, хранения, обработки и использования информации. Основные подходы к измерению количества информации.

2. ТЕОРИЯ КОДИРОВАНИЯ

Теория информации. Побуквенное кодирование. Разделимые коды. Префиксные коды. Критерий однозначности декодирования. Неравенство МакМиллана для делимых кодов. Оптимальные коды. Метод построения оптимальных кодов. Метод Хаффмана. Самокорректирующиеся коды. Коды Хэмминга. Коды Хэмминга, исправляющие одиночную ошибку.

3. ТЕОРИЯ АВТОМАТОВ

Конечные автоматы. Автоматные функции. Состояния автомата. Эквивалентность состояний. Теорема об эквивалентности состояний конечного автомата. Детерминированные функции. Задание детерминированных функций при помощи деревьев, вес функций. Ограничено-детерминированные функции. Схемы из логических элементов и элементов задержки. Реализация автоматных функций.

4. ГРАММАТИКИ

Классификация грамматик. Локализованные структуры. Канонические формы. Деревья. Двоичные деревья трансляции. Общий алгоритм. Алгоритм грамматического разбора. Диаграммы Вирта. Иерархия Хомского. Конечный автомат. Регулярные грамматики. Регулярные выражения. Регулярные события. Методы абстрактного синтеза. Структурная теория автоматов.

5. ТЕОРИЯ РАСПОЗНАВАНИЯ

Проблема распознавания. Общая характеристика задач распознавания и их типы. Математическая теория распознавания образов. Постановка задачи распознавания. Алгебраический подход к задаче распознавания. Геометрические процедуры распознавания. Алгоритмы распознавания, основанные на вычислении оценок. Структурные методы распознавания. Типы задач распознавания изображений. Распознавание и обработка изображений.

6. МАТЕМАТИЧЕСКАЯ КИБЕРНЕТИКА

Информация и управление. Математические аспекты кибернетики. Автоматическое регулирование. Программное управление и управление с обратной связью. Оптимальное управление. Методы прогнозирования. Теория принятия решений. Диалоговые системы оптимизации и имитации.

ТЕМЫ И ПЛАНЫ ЛАБОРАТОРНЫХ ЗАНЯТИЙ

Модуль 1. Теория информации

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

- определение количества информации;
- построение таблицы частот.

2. Разработка алгоритмов подсчета количества информации разными методами (2 часа):

– разработка алгоритма подсчета количества информации по Хартли и Шеннону на языке программирования.

3. Разработка программы формирования псевдослучайных чисел и оценка их качества статистическими тестами (2 часа):

- проверка равномерности распределения;
- проверка на статистическую независимость.

Модуль 2. Кодирование

4. Кодирование информации методами Шеннона–Фено и Хаффмана (2 часа):

- метод Шеннона–Фено;
- кодирование информации;
- метод Хаффмана.

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

- LZ-метод;
- алгоритм LZ78.

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

- обнаружение одиночной ошибки (2 метода).

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

Модуль 3. Автоматы

8. Комбинационные схемы (2 часа):

- построение схем из логических элементов;

- упрощение логических уравнений;
- построение субтрактора.

9. Регулярные выражения (2 часа):

- запись регулярных событий;
- применение методов абстрактного синтеза;
- работа с программой «Регулярные выражения».

Модуль 4. Грамматики

10. Диаграммы Вирта (2 часа):

- выделение синтаксических классов для описания языка

Паскаль;

- описание синтаксического класса «оператор».

11. Формальные грамматики. Разбор снизу вверх (алгоритм с возвратами) (2 часа):

- построение дерева разбора;
- алгоритм с возвратами.

Модуль 5. Распознавание образов.

Математическая кибернетика

12. Распознавание образов (2 часа):

- методы распознавания образов;
- работа с программой *Prog*.

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, Ъ с Ь, а также имеется знак пробела между словами, приведены в таблице 2.

Таблица 2

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, Ь 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_p – оказывается много меньше оптимальных сообщений – H_o . Допустим, нужно передать сообщение, содержащее количество информации, равное I . Источнику, обладающему энтропией на букву, равной H_p , придется затратить некоторое число n_p , то есть $I = n_p H_p$.

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

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

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

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

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.

Задания

Задание 1. Определить количество информации (по Хартли), содержащееся в заданном сообщении, при условии, что знаменателями 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 заданного 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 алгоритмов подсчета количества информации разными методами

Теоретические сведения

Теоретический материал находится на портале университета (Челябинский государственный педагогический университет (внутренний портал) > Учебно-методические материалы > inform > Bakalavriat > Inf > 4k > ТОИ).

Задания

Задание 1. Подготовить форму для подсчета количества информации методами Р. Хартли и К. Шеннона (см. рис. 1).

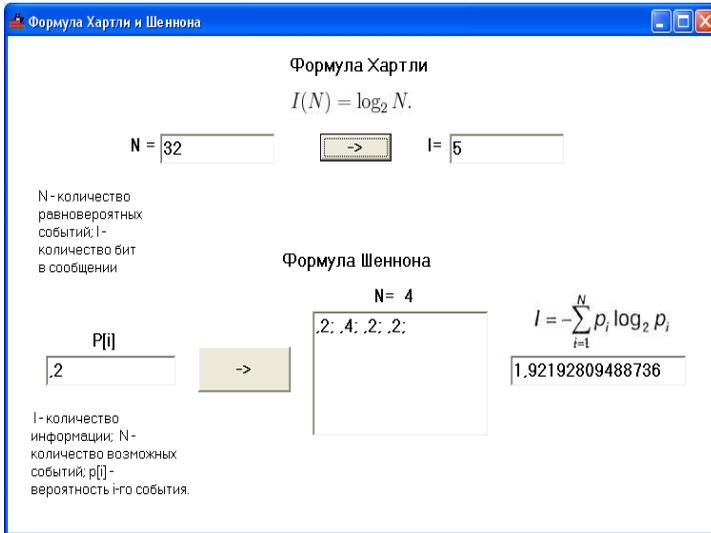


Рис. 1. Формула Хартли и Шеннона

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(label3.Caption)+1) ;
      memo1.Lines.Text:=memo1.Lines.Text+ edit1.Text+' '; ;
```

```
      edit2.Text:=floattostr(strtofloat(edit2.Text)+
                             (strtofloat(edit1.Text)*(-x))); ;
```

```
    end;
```

```
end;
```

```

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

```

Задания для самостоятельной работы

Рассмотреть метод вычисления количества информации, предложенный А.А. Харкевичем.

Лабораторная работа № 3

Разработка программы формирования псевдослучайных чисел и оценка их качества статистическими тестами

Теоретические сведения

1. 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

2. Частотный тест.

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. 2).

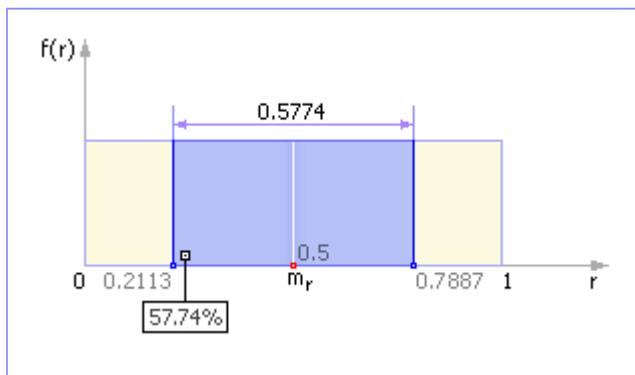


Рис. 2. Частотная диаграмма идеального ГСЧ в случае проверки его на частотный тест

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

Задания и примеры

Задание 1. Запустить программу ГСЧ.

Задание 2. Проверки на равномерность распределения.

Пример 1. Выполнение проверки по критерию «хи-квадрат».

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 г. 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. 3. Так как закон распределения эталонного ГСЧ равномерный, то (теоретическая) вероятность p_i попадания чисел в i -й 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).

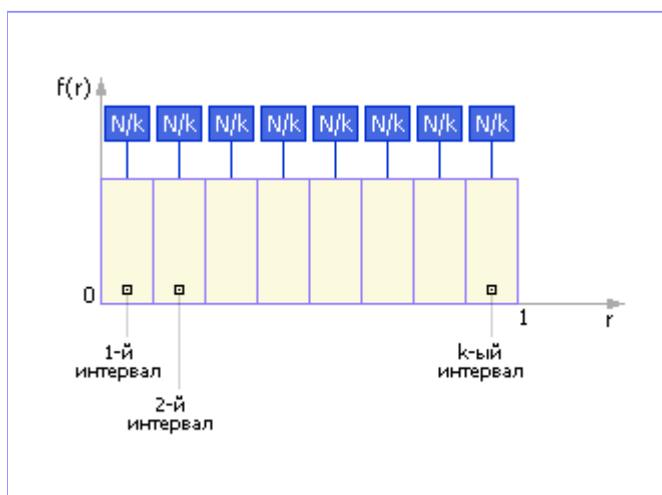


Рис. 3. Частотная диаграмма эталонного ГСЧ

Реальный ГСЧ будет выдавать числа, распределенные (причем, не обязательно равномерно!) по k интервалам и в каждый интервал попадет по n_i чисел (в сумме $n_1 + n_2 + \dots + n_k = N$). 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:

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

Из этой формулы следует, что чем меньше разность в каждом из слагаемых (а значит, и чем меньше значение $\chi^2_{\text{экс.}}$), 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 $p_i \cdot N$:

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

Наконец, запишем полученное выражение более компактно и упростим его:

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

Получили значение критерия «хи-квадрат» для *экспериментальных* данных. В табл. 3 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 $\chi^2_{\text{экс.}}$ будет меньше табулированного (теоретического) $\chi^2_{\text{теор.}}$ или равно ему.*

Если $\chi^2_{\text{экс.}}$ много больше $\chi^2_{\text{теор.}}$ (то есть p – велико), то генератор *не удовлетворяет* требованию равномерного распределения, так как наблюдаемые значения n_i слишком далеко уходят от теоретических $p_i \cdot N$ и не могут рассматриваться как случайные. Другими словами, устанавливается такой большой доверительный интервал, что ограничения на числа становятся очень

нежесткими, требования к числам – слабыми. При этом будет наблюдаться очень большая абсолютная погрешность.

Таблица 3

Некоторые процентные точки χ^2 -распределения

	p = 1%	p = 5%	p = 25%	p = 50%	p = 75%	p = 95%	p = 99%
v = 1	0,00016	0,00393	0,1015	0,4549	1,323	3,841	6,635
v = 2	0,02010	0,1026	0,5754	1,386	2,773	5,991	9,210
v = 3	0,1148	0,3518	1,213	2,366	4,108	7,815	11,34
v = 4	0,2971	0,7107	1,923	3,357	5,385	9,488	13,28
v = 5	0,5543	1,1455	2,675	4,351	6,626	11,07	15,09
v = 6	0,8721	1,635	3,455	5,348	7,841	12,59	16,81
v = 7	1,236	2,167	4,255	6,346	9,037	14,07	18,48
v = 8	1,646	2,733	5,071	7,344	10,22	15,51	20,09
v = 9	2,088	3,325	5,899	8,343	11,39	16,92	21,67
v = 10	2,558	3,940	6,737	9,342	12,55	18,31	23,21
v = 11	3,053	4,575	7,584	10,34	13,70	19,68	24,72
v = 12	3,571	5,256	8,438	11,34	14,85	21,03	26,22
v = 15	5,229	7,261	11,04	14,34	18,25	25,00	30,58
v = 20	8,260	10,85	15,45	19,34	23,83	31,41	37,57
v = 30	14,95	18,49	24,48	29,34	34,80	43,77	50,89
v = 50	29,71	34,76	42,94	49,33	56,33	67,50	76,15
v > 30	$v + \text{sqrt}(2v) \cdot x_p + 2/3 \cdot x_p^2 - 2/3 + O(1/\text{sqrt}(v))$						
$x_p =$	-2,33	-1,64	-0,674	0,00	0,674	1,64	2,33

Приемлемым считают p от 10% до 90%.

Еще Д. Кнут в своей книге «Искусство программирования» заметил, что иметь $\chi^2_{\text{эсп.}}$ маленьким тоже, в общем-то, нехорошо, хотя это и кажется, на первый взгляд, замечательно с точки зрения равномерности. Действительно, возьмите ряд чисел 0,1; 0,2; 0,3; 0,4; 0,5; 0,6; 0,7; 0,8; 0,9; 0,1; 0,2; 0,3; 0,4; 0,5; 0,6; ... – они идеальны с точки зрения равномерности, и $\chi^2_{\text{эсп.}}$

будет практически нулевым, но вряд ли вы их признаете случайными.

Если $\chi^2_{\text{эксп.}}$ много меньше $\chi^2_{\text{теор.}}$ (то есть p – мало), то генератор *не удовлетворяет* требованию случайного равномерного распределения, так как наблюдаемые значения n_i слишком близки к теоретическим $p_i \cdot N$ и не могут рассматриваться как случайные.

А вот 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.

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: $n_i, i = 1, \dots, k$.

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

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

где $p_i = 1/k$ – теоретическая вероятность попадания чисел в k -й интервал.

5. Путем сравнения экспериментально полученного значения $\chi^2_{\text{эксп.}}$ с теоретическим $\chi^2_{\text{теор.}}$ (из табл. 3) делается вывод о пригодности генератора для использования. Для этого:

а) входим в табл. 3 (строка = количество экспериментов – 1);
б) сравниваем вычисленное $\chi^2_{\text{эксп.}}$ с $\chi^2_{\text{теор.}}$, встречающимися в строке. При этом возможны три случая.

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

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

Третий случай: $\chi^2_{\text{эксп.}}$ лежит между значениями $\chi^2_{\text{теор.}}$ 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.

Пример 2. Выполнение проверки на частоту появления цифры в последовательности.

Случайное число 0,2463389991 состоит из цифр 2463389991, а число 0,5467766618 состоит из цифр 5467766618. Соединяя последовательности цифр, имеем: 24633899915467766618.

Понятно, что теоретическая вероятность p_i выпадения i -й цифры (от 0 до 9) равна 0.1.

Далее следует вычислить частоту появления каждой цифры в выпавшей экспериментальной последовательности. Например, цифра 1 выпала 2 раза из 20, а цифра 6 выпала 5 раз из 20.

Далее считают оценку и принимают решение по критерию «хи-квадрат».

Пример 3. Выполнение проверки появления серий из одинаковых цифр.

Обозначим через n_L число серий одинаковых подряд цифр длины L . Проверять надо все L от 1 до m , где m – это заданное пользователем число: максимально встречающееся число одинаковых цифр в серии.

В примере «24633899915467766618» обнаружены 2 серии длиной в 2 (33 и 77), то есть $n_2 = 2$ и 2 серии длиной в 3 (999 и 666), то есть $n_3 = 2$.

Вероятность появления серии длиной в L равна: $p_L = 9 \cdot 10^{-L}$ (теоретическая). То есть вероятность появления серии длиной в один символ равна: $p_1 = 0,9$ (теоретическая). Вероятность появления серии длиной в два символа равна: $p_2 = 0,09$ (теоретическая). Вероятность появления серии длиной в три символа равна: $p_3 = 0,009$ (теоретическая).

Например, вероятность появления серии длиной в один символ равна $p_L = 0,9$, так как всего может встретиться один символ из 10, а всего символов 9 (ноль не считается). А вероятность того, что подряд встретится два одинаковых символа «xx» равна $0,1 \cdot 0,1 \cdot 9$, то есть вероятность 0,1 того, что в первой позиции появится символ «x», умножается на вероятность 0,1 того, что во второй позиции появится такой же символ «x» и умножается на количество таких комбинаций 9.

Частотность появления серий подсчитывается по ранее разобранной формуле «хи-квадрат» с использованием значений p_L .

Примечание: генератор может быть проверен многократно, однако проверки не обладают свойством полноты и не гарантируют, что генератор выдает случайные числа. Например, генератор, выдающий последовательность 12345678912345..., при проверках будет считаться идеальным, что, очевидно, не совсем так.

Задание для самостоятельной работы

Выполнить проверку на статистическую независимость двух случайных чисел, полученных с помощью генератора случайных чисел (см. пример 2 и пример 3).

Индивидуальная работа № 1

Анализ текстов

Выполнить анализ текста на иностранном языке объемом не менее 500 слов (см. Приложение 4 и Приложение 5).

Вопросы к модулю 1

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. Шеннону?
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?

2. Закодировать фразу «Тише, мыши, тише, кот на крыше», используя метод Хаффмена.

Таблица 5

Символ	Кол-во	11	10	9	8	7	6	5	4	3	2	1
пробел	5	5	5	5	5	5	<u>6</u>	6	<u>8</u>	<u>10</u>	<u>12</u>	<u>18</u>
ш	4	4	4	4	4	<u>5</u>	5	<u>6</u>	6	8	10	12
е	3	3	3	3	<u>4</u>	4	5	5	6	6	8	
,	3	3	3	3	<u>3</u>	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	<u>3</u>	3	3	3					
к	2	2	2	2	3	3						
н	1	<u>2</u>	2	2	2							
о	1	1	2	2								
а	1	1	1									
м	1	1										
р	1											

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

Незакодированная фраза – 30*8 бит = 240 бит.

Закодированная фраза – 106 бит.

Задания

Использовать оба метода:

Задание 1. Закодировать фразу «наша саша шла по шоссе». Подсчитать количество бит, которое занимает закодированная фраза.

Задание 2. Закодировать фразу «во дворе трава на траве дрова». Подсчитать количество бит, которое занимает закодированная фраза.

Задание для самостоятельной работы

Закодировать фразу «сшит колпак да не по-колпаковски, надо колпак переколпаковать». Использовать метод Шеннона–Фено.

Индивидуальная работа № 2, 3

Кодирование методами Шеннона–Фено и Хаффмена

Закодировать фразу методами Шеннона–Фено и Хаффмена в соответствии с вариантом.

- | | |
|--|--|
| 1.
Ана, дэус, рики, паки,
Дормы кормы констунтаки,
Дэус дэус канадэус – бац! | 4.
Эне-бене, рики-таки,
Буль-буль-буль,
Караки-шмаки
Эус-деус-краснодеус 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. | 5.
Кони-кони, кони-кони,
Мы сидели на балконе,
Чай пили, воду пили,
По-турецки говорили. |
| 3.
Прибавь к ослиной голове
Еще одну, получишь две.
Но сколько б ни было ослов,
Они и двух не свяжут слов. | 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...

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.

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
Гурбаурбасентибряки.
Может – выйдет, может – нет,
З общем – полный Интернет

Лабораторная работа № 5

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 6. 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 $2^{K-1} - 1$.

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.

Таблица 6

Исходный текст	0.00.000. 01. 11. 111. 1111. 110. 0000.00000.1101. 1110.		
LZ-код	0.00.100.001.011.1011.1101. 1010.0110.10010.10001.10110.		
R	2	3	4
Вводимые коды	– 10 11 100 101 110 111 1000 1001 1010 1011 1100		

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.

Пример. Рассмотреть, как программа LZ78 читает каждый байт *abcabcabc* (см. табл. 7).

Таблица 7

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

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) $aba-abcabcdabcd$, используя алгоритм LZ78. Привести пример распаковки.

Самостоятельно:

а) использовать 4-х символьный алфавит (00, 01, 10, 11) для сжатия данных;

б) сжать последовательность «всем всем всем и каждому скажу».

Индивидуальная работа № 4

Сжатие данных по методу Лемпеля–Зива

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

Вариант 1

0100101010010000101

Вариант 3

1110100110111001101

Вариант 5

0100101010010000101

Вариант 7

0010100110010000001

Вариант 9

10101001101100111010

Вариант 11

010110110110100010001

Вариант 13

000101110110100111

Вариант 15

0100001000000100001

Вариант 17

0100100010010000101

Вариант 19

1110100110110001101

Вариант 2

0001000010101001101

Вариант 4

0001000010101001101

Вариант 6

10110111100110001101

Вариант 8

01011011001010101011

Вариант 10

000100101100100010001

Вариант 12

110101011001100001001

Вариант 14

101000100101010001011

Вариант 16

0100101010010000101

Вариант 18

0001010010101001101

Вариант 20

0001001010101001101

2. Закодировать следующую фразу, используя код LZ78:

Вариант 1

кукуркукурекурекун

Вариант 3

лорлоралоранранлоран

Вариант 5

какатанекатанекатата

Вариант 7

долделдолдилделдил

Вариант 2

упупапекапекаупуп

Вариант 4

пропронепронепрнепронас

Вариант 6

менменаменаменатеп

Вариант 8

sarsalsarsanlasanl

Вариант 9	Вариант 10
kloklonkolonklonkl	tertrektekertektrek
Вариант 11	Вариант 12
bigbonebigborebigbo	commercommecommerce
Вариант 13	Вариант 14
webwerbweberweberweb	porpoterpoterporter
Вариант 15	Вариант 16
mantopmentopomantomen	gorogopoterpoterter
Вариант 17	Вариант 18
webwerbweberweberweb	sionsinossionsinos
Вариант 19	Вариант 20
comconcomconacom	mantopmentopomantomen

Лабораторная работа № 6

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 кон-

трольных 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 8

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 9

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. 10).

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.

Таблица 10

Семизначный код Хемминга

Десятичное число	Простой двоичный код				Код Хемминга						
					к	к	к	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

Вывод. 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 $r = \text{ent}(d_{min} - 1) / 2$ и обнаруживаются ошибки кратности $d_{min} - 1$ (здесь ent означает «целая часть»). Так, при контроле на нечетность $d_{min} = 2$ и обнаруживаются одиноч-

ные 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. После инвертирования имеем 000. Теперь 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.

Задания

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

Лабораторная работа № 7

Защита реферата (Помехоустойчивое кодирование)

Задание для лабораторной работы № 7

ПРАКТИЧЕСКОЕ ЗАНЯТИЕ № 7

ЗАЩИТА РЕФЕРАТА (ПОМЕХОУСТОЙЧИВОЕ КОДИРОВАНИЕ)

ЗАДАНИЕ ДЛЯ ПРАКТИЧЕСКОГО ЗАНЯТИЯ № 7

1. Описать предложенный метод кодирования, привести 2 примера:

Вариант 1

инверсный

Вариант 3

БЧХ

Вариант 5

с пороговым декодированием

Вариант 7

Варшамова

Вариант 9

сверточный

Вариант 11

мажоритарный

Вариант 13

Вариант 2

Голея

Вариант 4

Грея

Вариант 6

Рида–Маллера

Вариант 8

цепной

Вариант 10

Плоткина

Вариант 12

Файра

Вариант 14

Абрамсона	Рида–Соломона
Вариант 15	Вариант 16
с постоянным весом	корреляционный
Вариант 17	Вариант 18
итеративный	с четным числом 1
Вариант 19	Вариант 20
Миласа–Абрамсона	по максимуму правдоподобия

2. Реализовать алгоритм (из задания 1) в среде Delphi (на 0,2–0,5 балла).

Вопросы к модулю 2

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 $\delta = (a \rightarrow 01, b \rightarrow 110, c \rightarrow 101)$.
Удовлетворяет ли данная схема неравенству МакМиллана?
12. Является ли разделимая схема префиксной?

13. Задача оптимального кодирования. Метод Шеннона–Фено.

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?

18. Как осуществляется передача дискретных сообщений по каналу с шумами.

19. В чем смысл помехоустойчивого кодирования?

20. Какие существуют помехоустойчивые коды?

21. Дать определение избыточности.

22. Перечислить виды избыточности.

23. Какие помехи вызывают внешние источники помех?

24. Что понимают под кодовым расстоянием?

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

26. Какие искажения называются краевыми? Дробления?

27. Какие коды называются непрерывными? Разделимыми?

28. В чем смысл кода Хемминга?

29. Какой код является оптимальным?

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

МОДУЛЬ 3. АВТОМАТЫ

Лабораторная работа № 8 Комбинационные схемы

Теоретические сведения

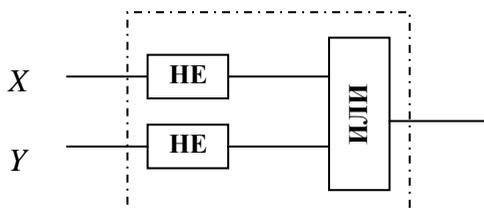
Теоретический материал находится на портале университета (<http://portal/Education/Forms/AllItems.aspx?RootFolder=%2fEducation%2finform%2fBakalavriat%2fIA%2f4k%2fd0%a2%2d0%9e%2d0%98&FolderCTID=&View=%7b33DF6B39%2d6239%2d47F5%2dB4B2%2d7C0ACA62E386%7d>).

Пример. Построить схему из логических элементов.

Упростим логическое уравнение:

$$\begin{aligned} X \cdot \bar{Y} \vee \bar{X} \cdot Y \vee \bar{X} \cdot \bar{Y} &= \bar{Y} \cdot (X \vee \bar{X}) \vee \bar{X} \cdot Y = \\ &= \bar{Y} \vee \bar{X} \cdot Y = \bar{Y} \vee \bar{X} \cdot \bar{Y} \vee Y = \bar{Y} \vee \bar{X} \end{aligned}$$

Схема будет выглядеть следующим образом:



Задания

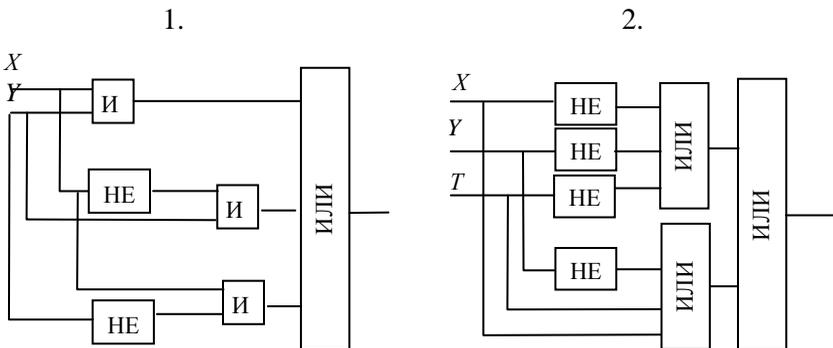
Задание 1. Построить схемы из логических элементов, предварительно упростив логические уравнения.

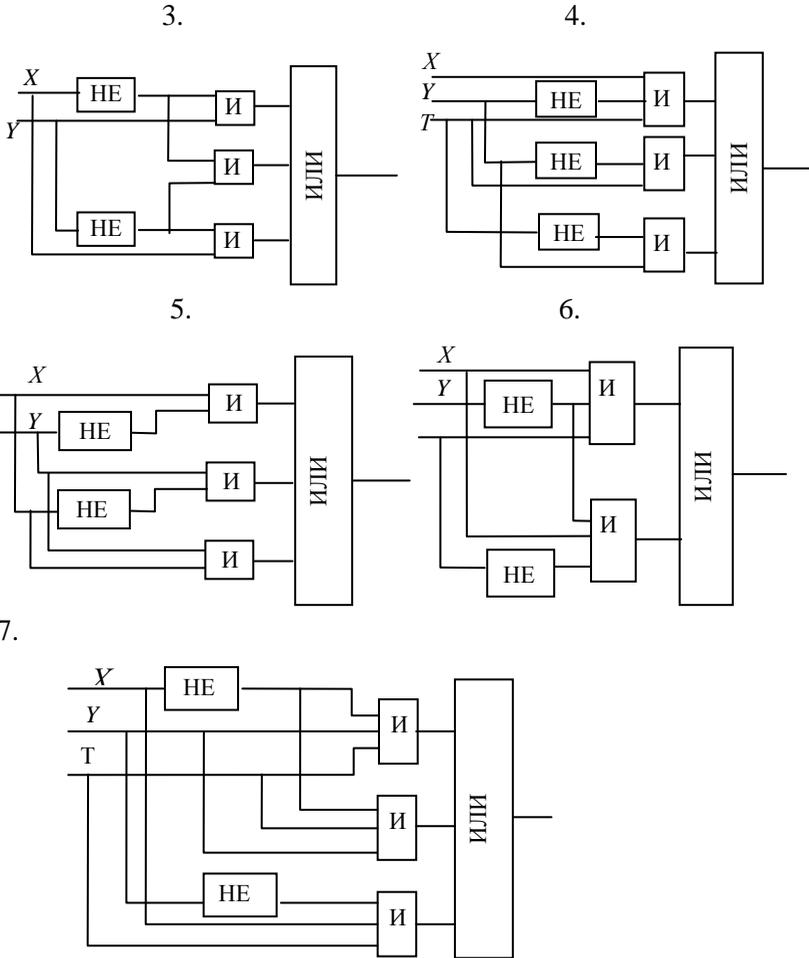
- 1) $\bar{X} \cdot \bar{Y} \vee X \cdot \bar{Y} \vee X \cdot Y$;
- 2) $X \cdot Y \cdot T \vee X \cdot Y \cdot \bar{T} \vee \bar{X} \cdot \bar{Y} \cdot \bar{T} \vee X \cdot \bar{Y} \cdot \bar{T}$;
- 3) $X \cdot Y \cdot \bar{T} \vee X \cdot \bar{Y} \cdot T \vee \bar{X} \cdot Y \cdot T \vee X \cdot Y \cdot T$;
- 4) $\bar{X} \cdot Y \cdot \bar{T} \vee \bar{X} \cdot \bar{Y} \cdot T \vee \bar{X} \cdot \bar{Y} \cdot \bar{T} \vee X \cdot Y \cdot \bar{T}$;
- 5) $P(X, Y, T) = X \cdot \bar{Y} \cdot T \vee X \cdot \bar{Y} \cdot \bar{T} \vee \bar{X} \cdot \bar{Y} \cdot \bar{T} \vee X \cdot \bar{Y}$.

Основные законы алгебры логики

Закон	Для ИЛИ	Для И
Переместительный	$x \vee y = y \vee x$	$x \cdot y = y \cdot x$
Сочетательный	$x \vee (y \vee z) = (x \vee y) \vee z$	$x \cdot (y \cdot z) = (x \cdot y) \cdot z$
Распределительный	$x \cdot (y \vee z) = x \cdot y \vee x \cdot z$	$x \vee (y \cdot z) = (x \vee y) \cdot (x \vee z)$
Правила де Моргана	$\overline{x \vee y} = \bar{x} \cdot \bar{y}$	$\overline{x \cdot y} = \bar{x} \vee \bar{y}$
Идемпотенции	$x \vee x = x$	$x \cdot x = x$
Поглощения	$x \vee (x \cdot y) = x$	$x \cdot (x \vee y) = x$
Склеивания	$(x \cdot y) \vee (\bar{x} \cdot y) = y$	$(x \vee y) \cdot (\bar{x} \vee y) = y$
Операция переменной с ее инверсией	$x \vee \bar{x} = 1$	$x \cdot \bar{x} = 0$
Операция с константами	$x \vee 0 = x, x \vee 1 = 1$	$x \cdot 1 = x, x \cdot 0 = 0$
Двойного отрицания	$\overline{\bar{x}} = x$	

Задание 2. По схеме восстановить функцию алгебры высказываний. Оптимизировать представленную схему.





Задание 3. Построить схему одnorазрядного субтрактора (от англ. *subtract* – вычитать)

Вычитание сводится к сложению **дополнительного кода вычитаемого**, т.е. отрицательного числа.

Получить дополнительный код можно следующим путем:

а) записать внутреннее представление **положительного** числа X ;

б) инвертируется двоичный код числа (заменяем 0 на 1 и 1 на 0);

в) к полученному числу добавляем единицу.

Пример.

Таблица 12

45 – 39	а) $(45)_{10} = (00101101)_2$ $(39)_{10} = (00100111)_2$
00101101	б) Инвертируем вычитаемое
–	11011000;
00100111	в) Находим сумму вычитаемое + 1
00000110	11011000
	+
	1
	11011001;
	Результат:
	00101101
	11011001
	1 00000110

Единица в 9-м разряде свидетельствует о том, что получено положительное число, $(00000110)_2 = (6)_{10}$.

Задание 4. Выполнить следующие действия: $17 - 7$; $5 - 10$.

Задание для самостоятельной работы

Построить схему субтрактора.

Лабораторная работа № 9

Регулярные выражения

Теоретические сведения

1. Представление событий в автоматах

Для описания автоматов на языке регулярных событий вводят ряд операций над событиями, т.е. строят алгебру собы-

тий. Рассмотрим алгебру событий, введенную Клини и усовершенствованную академиком В.М. Глушковым.

Алгебра событий включает три операции:

– дизъюнкцию (объединение) событий;

– произведение событий;

– итерацию событий.

Дизъюнкцией событий S_1, S_2, \dots, S_k называют событие $S = S_1 \vee S_2 \vee \dots \vee S_k$, состоящее из всех слов, входящих в события S_1, S_2, \dots, S_k .

Примеры и задания

Пример. Событие S_1 содержит слова x_1, x_2x_1, x_1x_1 , т.е. $S_1 = (x_1, x_2x_1, x_1x_1)$, а $S_2 = (x_2, x_1x_2)$. Тогда $S = S_1 \vee S_2 = (x_1, x_2, x_1x_1, x_1x_2, x_2x_1)$.

Задание 1.

а) $S_1 = (x_1, x_1x_1, x_2x_1)$, а $S_2 = (x_2, x_1x_2)$. Получить $S = S_1 \vee S_2$.

б) $S_1 = (x_2, x_1x_2, x_2x_1)$, а $S_2 = (x_1, x_2x_1)$. Получить $S = S_1 \vee S_2$.

Произведением событий S_1, S_2, \dots, S_k называется событие $S = S_1 * S_2 * \dots * S_k$, состоящее из всех слов, полученных приписыванием к каждому слову события S_1 каждого слова события S_2 , затем слова события S_3 и т.д.

Примеры и задания

Пример. S_1 и S_2 те же. $S = S_1 * S_2 = (x_1x_2, x_1x_1x_2, x_2x_1x_2, x_2x_1x_1x_2, x_1x_1x_2, x_1x_1x_1x_2)$.

Произведение событий не коммутативно, т.е. слова, входящие в события S_1S_2 и S_2S_1 различны, т.е. $S_1S_2 \neq S_2S_1$. Поскольку произведение не коммутативно, следует различать операции «умножение справа» и «умножение слева». Например, относительно произведения событий S_1S_2 можно сказать, что событие S_2 умножено на событие S_1 справа, а событие S_1 на S_2 слева.

Задание 2. S_1 и S_2 те же, что в а) и б). Получить $S = S_1 * S_2$.

Третьей операцией, применяемой в алгебре событий, является одноместная операция «итерация», которая применима только к одному событию. Для обозначения итерации вводят фигурные скобки, которые называются итерационными.

Итерацией события S называется событие $\{S\}$, состоящее из пустого слова e и всех слов вида S, SS, SSS и т.д. до бесконечности. Т.е. $\{S\} = e \vee S \vee SS \vee SSS \vee \dots$

Примеры и задания

Пример. $S = (x_2, x_1x_2)$.

$\{S\} = (e, x_2, x_2x_2, x_2x_2x_2, \dots, x_1x_2, x_1x_2x_1x_2, \dots, x_2x_1x_2, x_1x_2x_2, \dots)$.

Задание 3. $S = (x_1, x_2x_1)$. Получить $\{S\}$.

При синтезе конечных автоматов важнейшую роль играют регулярные события. Пусть дан конечный алфавит $X = (x_1, x_2, \dots, x_m)$.

Определение. Любое событие, которое можно получить из букв данного алфавита с помощью конечного числа операции дизъюнкции, произведения и итерации, называется **регулярным событием**, а выражение, составленное с помощью этих операций – **регулярным выражением**.

Рассмотрим пример составления регулярного выражения, определяющего закон функционирования конечного автомата.

Задание 4. Открыть программу *FA*, пример «Составления регулярного выражения для автомата сравнения чисел». Количество разрядов числа – произвольное.

Вопросы:

1. Как формируются выходные сигналы y_1, y_2, y_3 ?
2. Что поступает на входы автомата сравнения?
3. Какие буквы содержит выходной алфавит?
4. На пятом этапе ввести свои значения в соответствующие поля и просмотреть работу автомата.

Теоретические сведения

2. Методы абстрактного синтеза. Задача абстрактного синтеза заключается в составлении таблиц переходов и выходов автоматов по заданным условиям его функционирования, представленным в форме регулярных выражений.

Абстрактный синтез обычно выполняется в два этапа:

1. Первый этап заключается в получении таблиц переходов и выходов в некоторой исходной форме. Построенный по этим таблицам автомат обычно содержит «лишние» внутренние состояния.

2. На втором этапе производится минимизация количества внутренних состояний заданного автомата.

Синтезируемый автомат может быть задан либо как автомат Мура, либо как автомат Мили. Поскольку для автомата Мура всегда можно построить эквивалентный автомат Мили, то достаточно рассмотреть алгоритм синтеза автомата Мура, который проще автомата Мили.

Задание 5. Открыть программу *FA пример «Обучение. Абстрактный синтез конечных автоматов»*. Рассмотреть пример абстрактного синтеза автомата для случая, когда регулярные отношения составлены без применения операции итерации, а также с применением операции итерации.

1) $S1=X_1X_1X_2\vee X_2X_2|Y_1$, $S2=X_1X_2\vee X_1X_2|Y_2$. $S3=S1\vee S2$. Выполнить абстрактный синтез автомата.

2) $\{X_1X_1\vee X_1X_2\}X_1X_1\{X_2\}X_1$. Выполнить абстрактный синтез автомата с применением операции итерации.

Задание 6. Запустить систему *«Конечные автоматы и регулярные выражения»*. Добавить состояние или букву входного алфавита можно с помощью кнопок  и . На рис. 4 показан пример заполнения полей «Конечный алфавит» и «Регулярное выражение».

После того, как задан конечный алфавит и введено регулярное выражение, нажать кнопку . Установить входную цепочку можно по кнопке . Установка входной цепочки на ленту и автомата в начальное состояние осуществляется по кнопке .

Входная цепочка вводится в поле справа. В полях «Лента» и «Текущее состояние» отображается их содержимое на каждом шаге работы алгоритма. Переход на следующий шаг – .

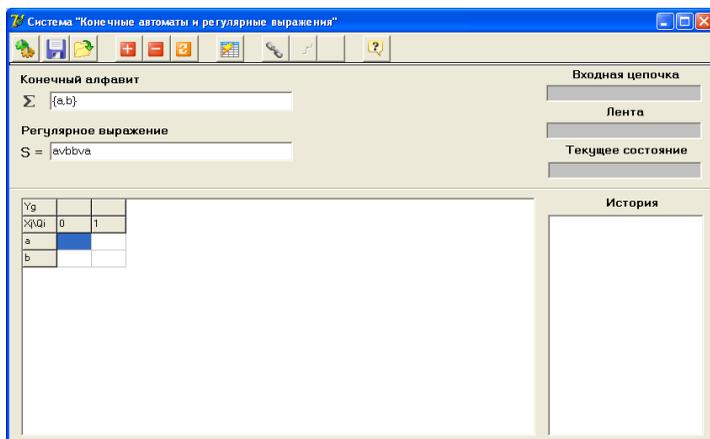


Рис. 4. Пример заполнения полей «Конечный автомат» и «Регулярное выражение»

На рис. 5 в окне «История» отображается ход исполнения алгоритма и вывод о допустимости или недопустимости цепочки.

Задание 7. Задать конечный алфавит $\{a, b, c\}$ и регулярное выражение a^ncavba . Просмотреть работу алгоритма для разных входных цепочек.

Задание для самостоятельной работы.

Привести два примера регулярных выражений.

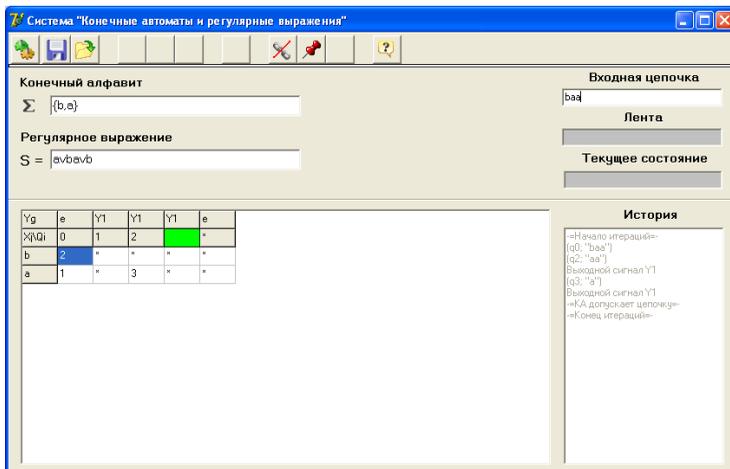


Рис. 5. Ход выполнения алгоритма и вывод о допустимости или недопустимости цепочки

Индивидуальная работа № 5

Автоматы. Комбинационные схемы

1. Упростить логическое выражение. Построить схему из логических элементов. $(X \supset Y = \bar{X} \vee Y)$.

Вариант 1

$$A \cdot B \vee \bar{A} \cdot B \vee A \cdot \bar{C} \vee C \cdot A \vee B \\ (\bar{A} \supset \bar{B}) \supset ((B \cdot C) \supset (A \cdot C))$$

Вариант 3

$$(X \vee Y) \cdot \bar{Z} \vee \bar{X} \cdot Z \vee Y \\ (C \supset A) \supset (\neg(B \vee C) \supset A)$$

Вариант 5

$$A \cdot B \vee ((A \vee B) \cdot (\bar{A} \vee \bar{B})) \\ (\neg(A \cdot (B \vee C)) \supset ((A \cdot B) \vee C))$$

Вариант 7

$$((A \cdot B) \vee (C \cdot B)) \cdot ((A \cdot B) \vee \bar{B}) \\ (((A \supset B) \supset \bar{A}) \supset \bar{B}) \supset \bar{C} \supset C$$

Вариант 2

$$(A \vee B) \cdot A \cdot (\bar{C} \vee (A \cdot C)) \\ ((A \supset B) \supset \bar{A}) \supset (A \supset (B \cdot A))$$

Вариант 4

$$(X \vee \bar{Y}) \cdot \bar{Z} \vee \bar{Y} \cdot (Z \vee Y) \\ \neg((A \cdot B) \supset A) \vee (A \cdot (B \vee C))$$

Вариант 6

$$A \cdot \bar{B} \vee ((\bar{A} \vee B) \cdot (\bar{A} \vee \bar{B})) \\ ((A \supset B) \supset (C \supset \bar{A})) \supset (\bar{B} \supset \bar{C})$$

Вариант 8

$$(((A \cdot B) \vee A) \vee C) \cdot (A \cdot \bar{B}) \\ (A \supset (B \supset C)) \supset ((A \supset \bar{C}) \supset (A \supset \bar{B}))$$

Вариант 9

$$(X \vee Y) \cdot \bar{Z} \vee \bar{X} \cdot Y \cdot Z \vee Y \cdot \bar{X} \\ (P \supset Q) \vee (P \supset (Q \cdot P))$$

Вариант 11

$$X \cdot Y \vee \bar{X} \cdot Y \vee X \cdot \bar{Z} \vee Z \cdot X \vee \bar{Y} \\ P \cdot (Q \supset P) \supset \bar{P}$$

Вариант 13

$$X \cdot Y \vee \bar{X} \vee X \cdot \bar{Z} \vee Z \vee \neg(X \vee \bar{Y}) \\ ((P \cdot (Q \vee \bar{P})) \cdot ((\bar{Q} \supset P) \vee Q))$$

Вариант 15

$$X \cdot Y \cdot Z \vee \bar{X} \cdot Y \cdot Z \vee X \cdot Y \cdot \bar{Z} \vee Z \cdot X \cdot \bar{Y} \\ ((X \supset (Y \cdot Z)) \supset (\bar{Y} \supset \bar{X})) \supset \bar{Y}$$

2. Выполнить следующие действия, построив дополнительный код (использовать 8-разрядный код):

Вариант 1

$$45 - 23; 12 - 34$$

Вариант 2

$$56 - 33; 23 - 44$$

Вариант 3

$$52 - 34; 21 - 33$$

Вариант 4

$$42 - 31; 14 - 55$$

Вариант 5

$$35 - 12; 22 - 45$$

Вариант 6

$$55 - 32; 32 - 44$$

Вариант 7

$$38 - 21; 25 - 32$$

Вариант 8

$$52 - 29; 19 - 42$$

Вариант 9

$$47 - 23; 22 - 45$$

Вариант 10

$$51 - 34; 15 - 28$$

Вариант 11

$$52 - 37; 25 - 67$$

Вариант 12

$$33 - 26; 23 - 50$$

Вариант 13

$$36 - 22; 24 - 51$$

Вариант 14

$$54 - 12; 25 - 55$$

Вариант 15

$$38 - 16; 30 - 47$$

Вариант 10

$$(Y \vee Z) \cdot \bar{Z} \vee \bar{X} \cdot (X \vee Y) \cdot \bar{Y} \\ \neg(P \supset \neg(Q \cdot P)) \supset (P \vee R)$$

Вариант 12

$$X \cdot Y \cdot Z \vee \bar{X} \cdot Y \cdot Z \vee X \cdot \bar{Y} \cdot \bar{Z} \vee Z \cdot X \\ ((P \cdot \bar{Q}) \supset Q) \supset (P \supset Q)$$

Вариант 14

$$Y \cdot Z \vee \bar{Z} \vee Y \cdot Z \vee \bar{Y} \vee \bar{Z} \cdot Y \\ ((Q \supset (P \cdot R)) \cdot \neg((P \vee R) \supset Q))$$

Вопросы к модулю 3

1. В чем заключается синтез автомата?
2. Для чего необходимо оптимизировать структуру автомата?
3. В каких единицах измеряется тактовая частота?
4. Какая логическая функция дает на выходе 0, только когда оба входа соответствуют 0?
5. В каком автомате функция Ψ определяет значение выходного сигнала только по одному аргументу?
6. Как называется функция, определяющая изменения состояний под воздействием входных сигналов?
7. Как задается автомат?
8. Что собой представляет функция переходов?
9. Что собой представляет функция выходов?
10. Как называется автомат, если функции переходов и выходов однозначно определены?
11. А чем отличие математической модели автомата Мили и схемы рекуррентных соотношений отличаются от математической модели и схемы рекуррентных соотношений абстрактного автомата?
12. Какой режим работы сети автоматов называют асинхронным?
13. Какой режим работы сети автоматов называют синхронным?
14. Что собой представляет автомат без памяти?
15. В чем заключается блочный синтез?
16. В чем заключается абстрактный синтез?
17. В чем заключается структурный синтез?

МОДУЛЬ 4. ГРАММАТИКИ

Лабораторная работа № 10

Диаграммы Вирта

Теоретические сведения

Наряду с текстовыми способами описания синтаксиса языков широко используются и графические метаязыки. Широко известность получил язык диаграмм Вирта, впервые примененный для описания синтаксиса языка Паскаль.

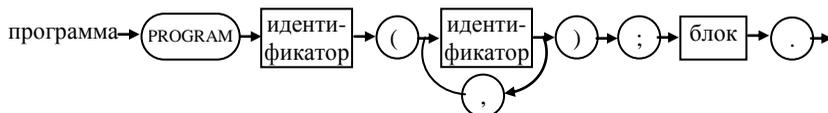
Метасимволы заменены следующими графическими обозначениями:

○ – терминальный символ, принадлежащий алфавиту языка;

○ (горизонтально вытянутое) – постоянная группа терминальных символов, определяющих название лексемы, ключевое слово и т.д.;

□ – нетерминальный символ, определяющий название правила.

Синтаксический класс **программа** выглядит следующим образом:



Задание.

1. Описать синтаксические классы: буква, цифра, идентификатор, целое без знака, число без знака;
2. Описать синтаксические классы: переменная, множитель, слагаемое, простое выражение, выражение;
3. Для синтаксического класса **оператор** описать: оператор присваивания, идентификатор процедуры, «операторные скобки» *begin ... end*, условный оператор, оператор присоединения и оператор безусловного перехода.

Синтаксический класс **блок**:

Индивидуальная работа № 6

Диаграммы Вирта

1. Построить диаграммы Вирта для следующих конструкций:

Вариант 1

Тип (set of ..., array), for

Вариант 3

Константа без знака, if

Вариант 5

Список полей,
идентификатор процедуры

Вариант 7

Переменная,
оператор присваивания

Вариант 9

Заголовок (процедуры или
функции), пустой оператор

Вариант 11

Идентификатор процедуры,
оператор вызова процедуры

Вариант 13

Тип (record, file), case

Вариант 2

Константа, while

Вариант 4

Список параметров, with

Вариант 6

Простой тип, repeat

Вариант 8

Множитель (константа, (),
идентификатор функции), goto

Вариант 10

Тип (file of ..., record),
оператор вызова функции

Вариант 12

Множитель (переменная,
not, []), метка

2. Привести пример грамматики (не менее 5 правил).

Лабораторная работа № 11

Формальные грамматики. Разбор снизу вверх (алгоритм с возвратами)

Теоретические сведения

1. Пусть $G_A = (V, W, P, \sigma)$ – грамматика,
где $V = \{A, B, C, \dots, Z\} \cup \{(), \times, /, +, -, \uparrow\}$

W = {<арифметическое выражение>, <операция типа умножения>, <терм>, <операция типа сложения>, <первичное>, <множитель>, <операнд> }

σ = <арифметическое выражение>

P = {<операция типа сложения> → + :1
 <операция типа сложения> → - :2
 <операция типа умножения> → × :3
 <операция типа умножения> → / :4
 <первичное> → <операнд> :5
 <первичное> → (<арифметическое выражение>) :6
 <множитель> → <первичное> :7
 <множитель> → <множитель> ↑ <первичное> :8
 <терм> → <множитель> :9
 <терм> → <терм> <операция типа умножения>
 <множитель> :10
 <арифметическое выражение> → <терм> :11
 <арифметическое выражение> → <операция типа
 сложения> <терм> :12
 <арифметическое выражение> → <арифметическое
 выражение> <операция типа сложения> <терм> :13
 <операнд> → A :14
 •
 •
 •
 <операнд> → Z :40}

Задания

Задание 1. Построить дерево разбора (для грамматики G_A):

- a) $a * b + c$;
- b) $b + ac$.

Задание 2. Запустить программы *Завершенность* и *Неудача*, предназначенные для демонстрации алгоритма грамматического разбора. Протестировать в пошаговом режиме.

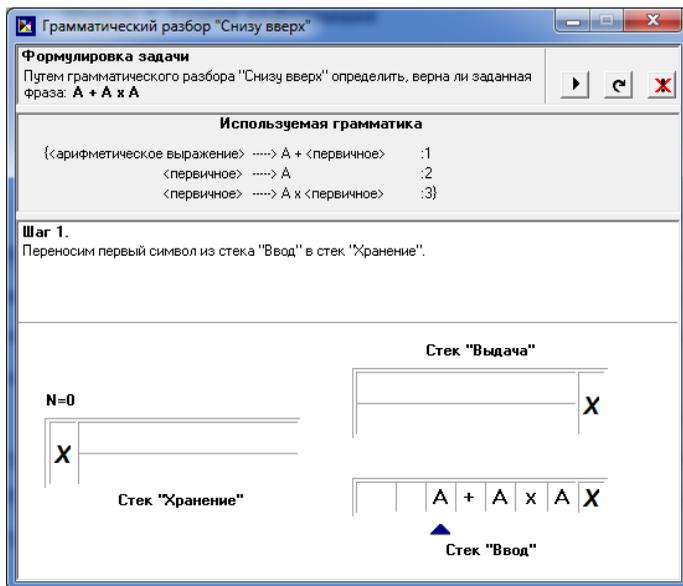


Рис. 6. Окно программы «Завершенность»

2. Алгоритм грамматического разбора.

В1. (выбрать следующий входной символ)

Если *Ввод* пуст, то перейти к В4;

Иначе выбрать из *Ввод* в *Хранение*;

Номер := 0.

В2. (поиск основы)

<искать правило подстановки с номером большим, чем Номер, аргумент которого совпадает с верхней подстрокой в стеке *Хранение*>

Если такого правила нет, то перейти к В2.

В3. (выдать возможную основу)

Поместить —| в стек *Выдача*;

Выбрать совпавшую подстроку из стека *Хранение* в стек *Выдача*;

Поместить корень и номер совпавшего правила в стек *Хранение*;

Номер:= 0;

Перейти к В2.

В4. (проверка завершенности)

Если стек *Хранение* содержит только σ /номер правила> на вершине, то поместить в стек *Выдача*. Выход – **успех**.

В5. (откат)

Выбирать из стека *Хранение* в стек *Ввод* до тех пор, пока пара <переменная><номер правила> не окажется на вершине стека *Хранение*;

Если нет, то Выход – **неудача**.

иначе выбрать из стека *Хранение*;

Номер := <номер правила>;

Выбирать из стека *Выдача* в стек *Хранение* до тех пор, пока не обнаружится символ —|.

Выбрать из стека *Выдача* символ —|;

Перейти к В2.

Задание 3. Продемонстрировать работу алгоритма на следующих примерах (для грамматики G_A):

а) $a + b$;

б*) bc / a .

3. Пусть $G_k = (T, N, P, \sigma)$ – грамматика,

где $T = \{if, then, :=, +, >, i, j\}$

$N = \{<инструкция>, <оператор>, <логическое выражение>, <простое выражение>, <переменная>\}$

$\sigma = <инструкция>$

$P = \{<переменная> \rightarrow i \quad :1$

$<переменная> \rightarrow j \quad :2$

$\langle \text{выражение} \rangle \rightarrow \langle \text{переменная} \rangle$:3
$\langle \text{выражение} \rangle \rightarrow \langle \text{выражение} \rangle + \langle \text{переменная} \rangle$:4
$\langle \text{логическое выражение} \rangle \rightarrow \langle \text{выражение} \rangle >$ $\langle \text{выражение} \rangle$:5
$\langle \text{оператор} \rangle \rightarrow \langle \text{переменная} \rangle := \langle \text{выражение} \rangle$:6
$\langle \text{оператор} \rangle \rightarrow \text{if } \langle \text{логическое выражение} \rangle$ $\text{then } \langle \text{оператор} \rangle$:7
$\langle \text{инструкция} \rangle \rightarrow \langle \text{оператор} \rangle$:8

Задание для самостоятельной работы

Представить дерево разбора для входной строки: $\text{if } j > i$
 $\text{then } j := i + i.$

Индивидуальная работа № 7

Формальные грамматики.

Алгоритм грамматического разбора

1. Построить дерево разбора (для грамматики G_A):

- | | |
|---------------------------------|------------------------------|
| 1) $a - b + c; cbb;$ | 16) $a + b - c; cab;$ |
| 2) $b - cc; b * c * b;$ | 17) $b - a + c; bc * b;$ |
| 3) $a - b + c; bc * b;$ | 18) $a - c - b; b / cb;$ |
| 4) $b - ca; b * c - b;$ | 19) $b - ca; b * c + b;$ |
| 5) $bc - ca; b * c + b;$ | 20) $b - c * a; bc + b;$ |
| 6) $a * b - c / a; b * a - a;$ | 21) $a * b - c; b * aa;$ |
| 7) $b / ca; b + a - b;$ | 22) $b - c * a; b * a - b;$ |
| 8) $b - c + a; b * c - b;$ | 23) $b / c + a; b - c * b;$ |
| 9) $b - c / ba; b - a * c - b;$ | 24) $b - c / b; a * c - b;$ |
| 10) $c * c / a; a ++ b;$ | 25) $c * c + a; a - bc;$ |
| 11) $a * + a; a + b + c;$ | 26) $c * ab; a - b - c;$ |
| 12) $c * cba; a / b - c;$ | 27) $c * b - a; b / a - c;$ |
| 13) $c * b - a; c * ba;$ | 28) $a + c / b; cb / a;$ |
| 14) $a / c + a; aa * b;$ | 29) $a / c - a; b * cc;$ |
| 15) $b * a + c; c ++ a - b;$ | 30) $b + c * a; c -- a - b.$ |

2. Продемонстрировать работу алгоритма трансляции на следующих примерах:

- | | |
|-------------------|-------------------|
| 1) $a - b * c$; | 16) cab ; |
| 2) $b - c * a$ | 17) $b - a + c$; |
| 3) $c * b - c$; | 18) a / cb ; |
| 4) $b - c * a$; | 19) $b - ca$; |
| 5) $c - c / a$; | 20) $a + cb$; |
| 6) $c / a + a$; | 21) $b * aa$; |
| 7) $b / c / a$; | 22) $ca - b$; |
| 8) $b - c + a$; | 23) $c * b - a$; |
| 9) $c - b + a$; | 24) $c - b / a$; |
| 10) $b * c / a$; | 25) $a / b - a$; |
| 11) $a + c * c$; | 26) $b - c * a$; |
| 12) $c * c - a$; | 27) $b - a - c$; |
| 13) $a * c - b$; | 28) $cc - a$; |
| 14) $a - a * b$; | 29) $ac - a$; |
| 15) $a - bb$; | 30) $b + c * a$. |

Вопросы к модулю 4

1. Что является предметом семиотики как науки?
2. Дать определение синтаксиса, семантики и прагматики.
3. Что собой представляет правило подстановки, строка терминальных символов, сентенциальная форма?
4. Как задается грамматика?
5. Привести пример грамматики.
6. Дать определение предложения.
7. Какие символы относятся к терминальным символам?
8. Что собой представляет синтаксическая переменная?
9. Может ли быть несколько правил подстановки с одинаковыми левыми частями?
10. Какая грамматика называется контекстно-свободной грамматикой?

11. Какая грамматика называется контекстно-зависимой грамматикой?
12. Что такое язык? Сентенциальная форма?
13. В чем заключается свойство локальности?
14. Любая последовательность правил может быть приведена к канонической форме?
15. Каноническая форма однозначна?
16. Какая цепочка называется левосторонней цепочкой двоичного дерева?
17. Какая цепочка называется максимальной цепочкой двоичного дерева?
18. Привести иерархию Хомского.
19. Как описывается недетерминированный конечный автомат?
20. Можно ли преобразовать недетерминированный конечный автомат в эквивалентный детерминированный конечный автомат?
21. Какой автомат называется детерминированным?
22. Как осуществляется перенос символа из входной цепочки в магазин?
23. Что собой представляет сдвиг, свертка, рекурсивный спуск?
24. Допускает ли класс контекстно-свободных языков распознавание с помощью недетерминированного конечного автомата со стековой (или магазинной) памятью.
25. Для чего используются стек выдачи, стек хранения, блок сравнения?
26. Какая грамматика называется обратимой, приведенной, операторной?

МОДУЛЬ 5. РАСПОЗНАВАНИЕ ОБРАЗОВ. МАТЕМАТИЧЕСКАЯ КИБЕРНЕТИКА

Лабораторная работа № 12 Распознавание образов

Теоретические сведения

Распознаванием образов называются задачи построения и применения формальных операций над числовыми или символическими отображениями объектов реального или идеального мира, результаты решения которых отражают отношения эквивалентности между этими объектами.

Распознавание образов (объектов, сигналов, ситуаций, явлений или процессов) – задача идентификации объекта или определения каких-либо его свойств по его изображению (оптическое распознавание) или аудиозаписи (акустическое распознавание) и другим характеристикам.

Образ – классификационная группировка в системе классификации, объединяющая (выделяющая) определенную группу объектов по некоторому признаку. Методика отнесения элемента к какому-либо образу называется решающим правилом. Метрика – способ определения расстояния между элементами универсального множества. Обучением обычно называют процесс выработки в некоторой системе той или иной реакции на группы внешних идентичных сигналов путем многократного воздействия на систему внешней корректировки. Самообучение отличается от обучения тем, что здесь дополнительная информация о верности реакции системе не сообщается.

Примеры задач распознавания образов: распознавание букв, распознавание штрих-кодов, распознавание автомобильных номеров, распознавание лиц и других биометрических данных, распознавание речи.

Задания

Задание 1. Запустить программу *Prog*. Познакомиться с вкладками Распознавание букв и цифр, введенных пользователем; Распознавание изображений; Распознавание и обучение.

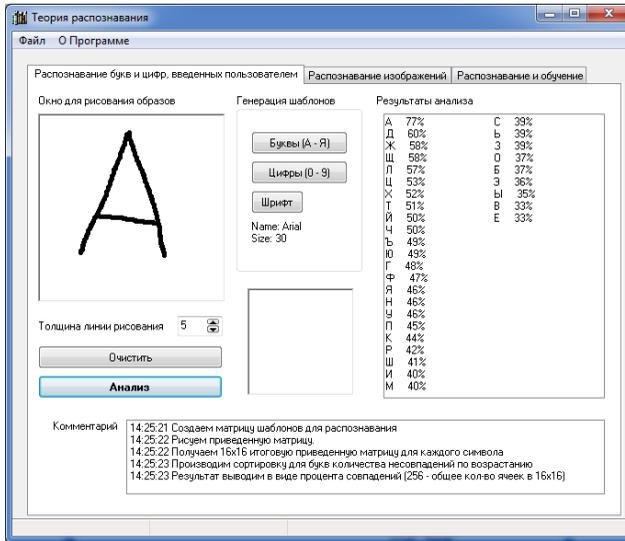


Рис. 7. Распознавание букв и цифр, введенных пользователем

Задание 2. Распознать буквы А, Д, Я и цифры 1, 4, 8 на вкладке Распознавание букв и цифр, введенных пользователем. Определить процент соответствия (см. рис. 7).

Задание 3. Обучить систему распознавать цифры 2, 5, 9 на вкладке Распознавание и обучение (см. рис. 8).

Пример. Распознавание почтового индекса.

Признаки для распознавания: x_1 – количество вертикальных линий минимального размера; x_2 – количество горизонтальных линий; x_3 – количество наклонных линий; x_4 – количество горизонтальных линий снизу объекта.

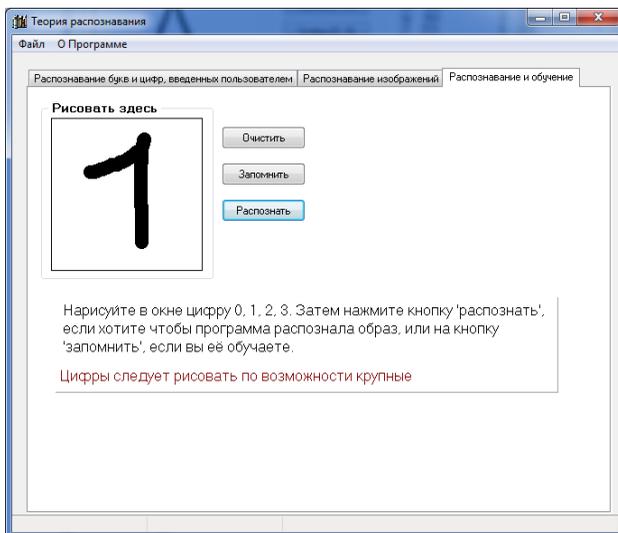


Рис. 8. Распознавание и обучение

Задание 4. Заполнить таблицу соответствия признакам для цифр индекса.

Таблица 13

Таблица соответствия признакам

x_1	x_2	x_3	x_4	цифры
4	2	0		0
				1
				2
				3
				4
				5
				6
				7
				8
				9
$\epsilon = 1$	$\epsilon = 1$	$\epsilon = 1$	$\epsilon = 1$	$\delta = 1$

Пример: Задача поиска геометрических аналогий.

Среди фигур второго ряда, представленных на рис. 9, требуется выбрать $X \in \{1, 2, 3, 4, 5\}$ такое, что A так соотносится с B , как C соотносится с X , и такое, которое лучше всего при этом подходит. В данном примере $X = 3$.

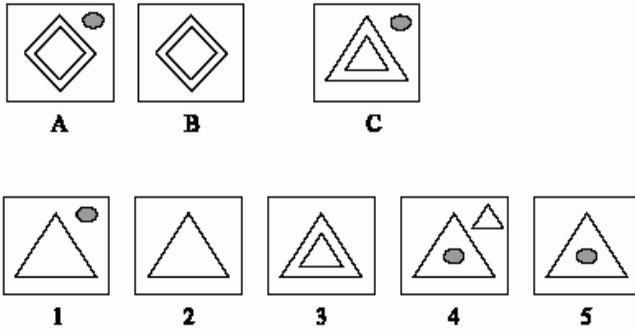


Рис. 9.

Задание 5. Сформулировать задачу поиска геометрических аналогий самостоятельно.

Пример: Правила преобразования в задаче поиска геометрических аналогий (рис. 10):

Правило 1 (исходное изображение): k выше m , k выше n , n внутри m .

Правило 2 (результатирующее изображение): n слева m .

Правило 3 (масштабирование, повороты):

- k исчезло;
- m изменение масштаба 1:1, вращение 00;
- n изменение масштаба 1:2, вращение 00.

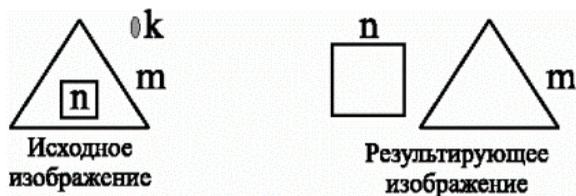


Рис. 10

Задание для самостоятельной работы

1. Среди фигур, представленных на рис. 11, требуется выбрать $X \in \{1, 2, 3\}$ такое, что A так соотносится с B , как C соотносится с X , и такое, которое лучше всего при этом подходит.

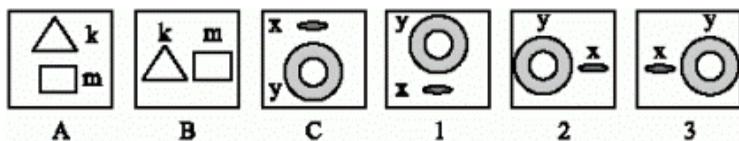


Рис. 11

2. Используя рис. 9, сформулировать задание на распознавание образов по аналогии с заданием 1 для самостоятельной работы.

Вопросы к модулю 5

1. Какие понятие в теории распознавания образов являются базовыми?
2. В чем состоит задача распознавание образов?
3. Что называется решающим правилом?
4. Что такое метрика?
5. Дать определение ситуации, состояния.
6. Что используется для распознавания сложных процессов и явлений?
7. Какие задачи относятся к задачам распознавания?
8. Дать характеристику задач распознавания.

9. В чем заключается алгоритм распознавания, основанный на вычислении оценок?

10. К какому этапу относится оценка погрешности в общей схеме преобразования данных?

11. Перечислить методы распознавания образов.

12. Какие методы распознавания образов относятся к интенсиональным?

13. Какие методы распознавания образов относятся к экстенсиональным?

14. Дать определение нейронной сети.

15. Обучение без учителя.

16. Основные понятия теории управления.

17. Управление с обратной связью.

18. Теория принятия решений (критерии).

19. Прогнозирование: основные понятия.

20. Оптимальное управление (примеры).

МОДУЛЬ 3. АВТОМАТЫ

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 $\Delta t = \frac{1}{2f_c}$, где f_c – верхняя граничная частота спектра непре-

рывного 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) $\frac{1}{2}$;
- 2) 1;
- 3) 2.

9. Поставить в соответствие подходу его характеристику:

- 1) структурный;
- 2) статистический;
- 3) прагматический;
- 4) семантический.

а) учет вероятностных характеристик источника сообщений;

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

- с) опирается на тезаурусный подход;
- д) определение меры количества информации;
- е) определение информационной емкости сообщения.

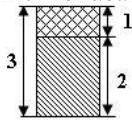
10. Опыт γ состоит в выполнении опытов α и β . Поставить в соответствие:

- | | |
|-------------------------|-----------------------------|
| 1) H_γ ; | а) условная энтропия; |
| 2) H_α ; | б) безусловная энтропия; |
| 3) $I_{\alpha,\beta}$. | с) энтропия сложного опыта; |
| | д) 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) $H_\gamma \leq H_\alpha + H_\beta$.

12. На рисунке показана зависимость энтропии и количества информации



. Поставить в соответствие:

- | | |
|-------|-------------------------|
| 1) 1; | е) H_β ; |
| 2) 2; | ф) $H_{\beta/\alpha}$; |
| 3) 3. | г) $H_{\beta/\beta}$; |
| | д) $I_{\alpha,\beta}$. |

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 инфор-

мации 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 $\beta_{i1}\beta_{i2}\dots\beta_{ik} = \beta_{j1}\beta_{j2}\dots\beta_{jk}$ следует, что $k = 1, \forall t \in \{1, 2, \dots, k\}$ и $i_t = j_t$, то схема кодирования называется:

- 1) префиксной;
- 2) постфиксной;
- 3) делимой.

19. Префиксная схема кодирования является делимой?

- 1) да;
- 2) нет.

20. Пусть $A = \{a, b\}$, $B = \{0, 1\}$ и $\delta = \langle a \rightarrow 010, b \rightarrow 01 \rangle$.

$l_\delta(P) = \sum_{i=1}^n P l_i$. 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 Фено строит схему кодирования:

- 1) оптимальную префиксную;
- 2) делимую префиксную;
- 3) оптимальную постфиксную.

22. Распределение вероятностей и код Хаффмена представлены в таблице. Чему равно $l_{\delta}(P)$?

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

- 1) 2,56; 2) 2.76; 3) 2,94; 4) 3.

23. Поставить в соответствие.

- | | |
|------------------|--------------------------------------|
| 1) C бит/с; | a) пропускная способность канала; |
| 2) H бит/букв; | b) энтропия источника информации; |
| 3) V букв/с. | 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?

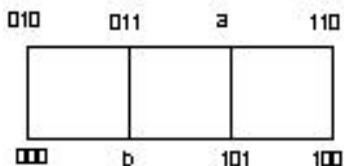
- 1) 2500 букв/с;
- 2) 5000 букв/с;
- 3) 5714 букв/с.

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 $\delta = \langle c \rightarrow 01, d \rightarrow 0 \rangle$. Схема δ :

- 1) разделима;
- 2) неразделимая;
- 3) префиксная.

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

- 1) да;
- 2) нет.

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

- 1) $\frac{1}{2}$;
- 2) $\frac{3}{8}$;
- 3) $\frac{5}{8}$;
- 4) 1.

32. Является ли префиксная схема неразделимой?

- 1) да;
- 2) нет.

33. Выбрать последовательность операций, записанных в порядке уменьшения их приоритета:

- 1) \neg , \cdot , $\&$;
- 2) \neg , $\&$, $;$
- 3) $\&$, \neg , $;$
- 4) $\&$, \cdot , \neg .

34. Двоичные цифры или булевы величины представляются в виде:

- 1) непрерывных кодов;
- 2) физических сигналов;
- 3) последовательностей сигналов.

35. Какая логическая функция дает на выходе 1, только когда оба входа соответствуют 1?

- 1) дизъюнкция;
- 2) конъюнкция;
- 3) импликация.

36. Какая логическая функция дает на выходе 0, только когда оба входа соответствуют 0?

- 1) дизъюнкция;
- 2) конъюнкция;
- 3) логическое следование.

37. Комбинационная схема может обучаться?

- 1) да;
- 2) нет.

38. Тактовая частота измеряется в:

- 1) байтах;
- 2) бодах;
- 3) бит/с;
- 4) герцах;
- 5) кулонах.

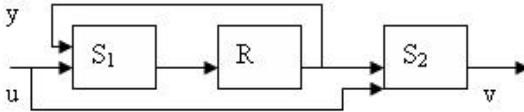
39. Конечное множество упорядоченных элементов называется:

- 1) регистром;
- 2) коммутатором;

3) двоичными элементами;

4) нет верного ответа.

40. На рисунке представлена схема. Поставить в соответствие.



1) S_1 ;

a) комбинационная схема;

2) R ;

b) регистр памяти;

c) последовательная схема;

d) нет верного ответа.

41. Автоматом называется система $S = \langle A, Q, V, \delta, \lambda \rangle$. Поставить в соответствие:

1) A ;

a) алфавит сообщений;

2) Q ;

b) алфавит состояний;

3) V ;

c) входной алфавит;

4) δ ;

d) выходной алфавит;

5) λ .

e) функция входов;

f) функция переходов;

g) функция выходов.

42. Автомат можно рассматривать как алгоритм с конечной памятью?

1) да;

2) нет.

43. Поставить в соответствие.

1) синтаксис – это... a) правила написания в данном языке;

2) семантика – это... b) правила истолкования языковых конструкций;

c) правила, определяющие пользовательскую сторону языка.

44. Строки терминальных символов называются:

1) предложениями;

2) терминальными символами;

3) синтаксическими переменными.

45. В данной подстановке $\xi \rightarrow \bar{x}$ ξ является:

- 1) предложением языка;
- 2) синтаксической переменной;
- 3) строкой терминальных символов.

46. Дана грамматика $G = (\Sigma, T, \Pi, \delta)$. Поставить в соответствие:

- | | |
|---------------|--|
| 1) Σ ; | a) правило подстановки; |
| 2) T ; | b) подстрока подстановки; |
| 3) Π ; | c) начальный символ грамматики; |
| 4) δ . | d) множество переменных символов; |
| | e) множество терминальных символов; |
| | f) пронумерованное множество правил подстановок. |

47. Множество (необязательно конечное) строк терминальных символов в соответствии с правилами грамматики – это:

- 1) сентенциальная форма;
- 2) предложение;
- 3) язык.

48. Если $G = (\Sigma, T, \Pi, \delta)$ есть грамматика с фразовой структурой и каждое правило $\bar{x} \rightarrow \bar{y}$ из G имеет вид $\bar{x} = \bar{u}_1 \xi \bar{u}_2$, $\bar{y} = \bar{u}_1 Z \bar{u}_2$, где $Z \neq \varepsilon$, то G называется:

- 1) контекстно-свободной грамматикой;
- 2) контекстно-зависимой грамматикой.

49. Любая последовательность правил может быть приведена к канонической форме?

- 1) да;
- 2) нет.

50. Каноническая форма однозначна?

- 1) да;
- 2) нет.

51. Даны цепочки a_1, \dots, a_n и b_1, \dots, b_m . Если существует такое $j \geq 0$, что $a_i = b_{j+i}$, $i = 1, \dots, n$, то

- 1) цепочка b_1, \dots, b_m является максимальной;
- 2) цепочка a_1, \dots, a_n содержится в цепочке b_1, \dots, b_m ;

3) последовательность a_1, \dots, a_n называется левосторонней цепочкой двоичного дерева.

52. При обнаружении возможной основы при грамматическом разборе эта основа передается:

- 1) в стек хранения;
- 2) в стек выдачи;
- 3) в блок сравнения.

53. Ввод – обычная строка переменных, которые организованы как:

- 1) очередь;
- 2) дерево;
- 3) стек.

54. Дана грамматика $G = (T, W, P, I)$, описывающая язык $\{a^{2k-1}\}$; $T = \{a\}$, $W = \{I\}$, $P: \begin{cases} I \rightarrow a \\ I \rightarrow Iaa \end{cases}$. Является ли последовательность $aaaa$ предложением на данном языке?

- 1) да;
- 2) нет.

55. Базовым в теории распознавания является понятие:

- 1) алгебры;
- 2) матрицы;
- 3) множества;
- 4) многочлена.

56. Методика отнесения элемента к какому-либо образу называется:

- 1) метрикой;
- 2) распознаванием;
- 3) решающим правилом.

57. Состояние – это:

1) некоторая совокупность состояний сложного объекта, каждая из которых характеризуется одними и теми же или схожими характеристиками объекта;

2) определенная форма отображения измеряемых текущих (или мгновенных) характеристик наблюдаемого объекта;

3) каждое отображение какого-либо объекта на воспринимающие органы распознающей системы, независимо от его положения относительно этих органов.

58. Некоторую совокупность состояний сложного объекта, каждая из которых характеризуется одними и теми же или схожими характеристиками объекта, принято назвать:

- 1) изображением;
- 2) ситуацией;
- 3) состоянием.

59. Для распознавания зрительных и слуховых образов используются:

- 1) состояние и ситуации;
- 2) изображения объекта и образы;
- 3) нет верного ответа.

60. Среди перечисленных задач одна не относится к задачам распознавания:

1) для этих задач трудно строить формальные теории и применять классические математические методы;

2) в этих задачах возможна «плохая» информация (информация с пропусками, разнородная и др.);

3) в этих задачах нельзя вводить понятие аналогии или подобия объектов.

61. Поставить в соответствие типу задачи распознавания ее характеристику:

- 1) задача распознавания;
- 2) задача автоматической классификации.

а) отнесение предъявленного объекта по его описанию к одному из заданных классов (обучение с учителем);

б) разбиение множества объектов (ситуаций) по их описаниям на систему непересекающихся классов (таксономия, кластерный анализ, обучение без учителя);

с) задача, в которой решение должно относиться к некоторому моменту в будущем.

62. Восстановить алгоритм распознавания, основанный на вычислении оценок:

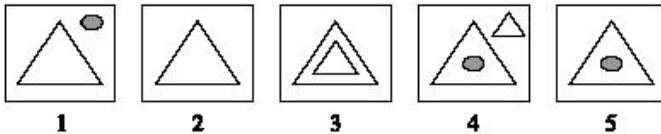
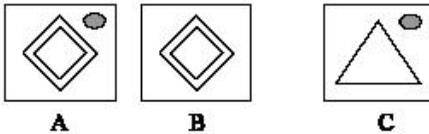
1) удалим произвольный набор признаков из строк $\omega_1, \omega_2, \dots, \omega_m$;

2) обозначим полученные строки через $S\omega_1, S\omega_2, \dots, S\omega_m, S\omega'$;

3) зададим полный набор признаков x_1, \dots, x_n ;

4) Выделим систему подмножеств множества признаков S_1, \dots, S_k .

63. Среди фигур второго ряда, представленных на рисунке, требуется выбрать $X \in \{1, 2, 3, 4, 5\}$ такое, что A так соотносится с B , как C соотносится с X , и такое, которое лучше всего при этом подходит.



1) $X = 2$;

2) $X = 3$;

3) $X = 5$.

64. В общей схеме преобразования данных оценка погрешности относится к этапу:

1) корректировки;

2) обработки.

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 курсов «Теоретические основы информатики», «Теория информации» и «Теоретические основы информатики и современных информационных технологий».

Для организации самостоятельной работы студентов, обучающихся по направлениям «Педагогическое образование» и «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 дисциплин «Теоретические основы информатики» и «Теория информации» используется модульно-рейтинговая система контроля знаний студентов. В приложении приведен пример оценки сформированности компетенций по указанным дисциплинам.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Безручко В.Т. Информатика: (курс лекций): учеб. пособие для вузов / В.Т. Безручко. – М.: ИНФРА-М: ФОРУМ, 2009.
2. Велихов А.В. Основы информатики и компьютерной техники: учеб. пособие для вузов / А. Велихов. – М.: СОЛОН Пресс, 2007.
3. Гагарина Л.Г. Современные проблемы информатики и вычислительной техники: учеб. пособие для подготовки магистров / Л.Г. Гагарина, А.А. Петров. – М.: Форум: Инфра-М, 2011.
4. Информатика: базовый курс: учеб. пособие для техн. вузов / ред. С.В. Симонович. – 3-е изд. – СПб.: Питер, 2012.
5. Информатика и информационные технологии: учеб. пособие для вузов / ред. Ю.Д. Романова. – 4-е изд., перераб. и доп. – М.: Эксмо, 2010.
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. – 3-е изд., стер. – М.: Омега-Л, 2007.
9. Мельников В.П. Информационные технологии: учеб. для вузов / В.П. Мельников. – 2-е изд., стер. – М.: Академия, 2009.
10. Подчукаев В.А. Теория информационных процессов и систем: учеб. пособие для вузов / В.А. Подчукаев. – М.: Гардарики, 2007.
11. Распознавание образов и машинное восприятие / А.С. Потапов. – СПб.: Политехника, 2007.
12. Тропченко А.Ю. Методы сжатия изображений, аудиосигналов и видео / А.Ю. Тропченко, А.А. Тропченко // Теоретическая информатика. – СПб.: СПбГУ ИТМО, 2009. – 108 с.
13. Информационно-коммуникационные технологии в образовании. – Режим доступа: URL: <http://www.ict.edu.ru/lib/> (дата обращения 16.01.2012).
14. Лидовский В.В. Теория информации: уч. пособие / В.В. Лидовский. – М.: Компания Спутник+, 2004. – 111 с.
15. Теоретическая информатика. – Режим доступа: URL: http://se.hse.ru/11003945/teor_inf (дата обращения 9.08.2015).

ПРИЛОЖЕНИЯ

Приложение 1

Рабочая (модульная) программа

Модуль 1. Теория информации		
Цель – знать основные понятия курса и основные подходы к измерению количества информации		
<i>Содержание раздела</i>		
<i>План лекций</i>	<i>План лабораторных работ</i>	<i>План семинарских занятий</i>
Лекция № 1. Роль информации в современном обществе. 1. Понятие об информации. 2. Различные определения информации. 3. Формы представления информации. Лекция № 2. Измерение количества информации. 1. Метод Хартли. 2. Статистический подход. 3. Энтропия. 4. Семантический подход. 5. Прагматический подход. Лекция № 3. Энтропия. 1. Безусловная энтропия. 2. Энтропия сложных событий.	1. Вычисление статистических характеристик текстовой информации (2 часа): – определение количества информации; – построение таблицы частот. 2. Разработка алгоритмов подсчета количества информации разными методами (2 часа): – разработка алгоритма подсчета количества информации по Хартли и Шеннону на ЯП.	Не предусмотрены учебным планом

<i>План лекций</i>	<i>План лабораторных работ</i>	<i>План семинарских занятий</i>
<p>Лекция № 4. Передача дискретных сообщений по каналу без шумов и с шумами.</p> <p>1. Первая теорема Шеннона. Обратная теорема Шеннона.</p> <p>2. Вторая теорема Шеннона.</p> <p>3. Помехи и борьба с ними.</p> <p>Методическое обеспечение лекций</p> <ul style="list-style-type: none"> • Презентации лекций хранятся на внутреннем портале университета. • Список литературы [1, 2, 6, 14, 15, 17] 	<p>3. Разработка программы формирования псевдослучайных чисел и оценка их качества статистическими тестами (2 часа):</p> <ul style="list-style-type: none"> – проверка равномерности распределения; – проверка на статистическую независимость 	<p>Не предусмотрены учебным планом</p>
<i>Самостоятельная работа</i>		
Инвариантная часть	Вариативная часть	
	Сдача индивидуальных заданий досрочно +0,2	
Контроль (итоговый тест)		
<i>Паспорт оценочных средств по разделу</i>		
Код контролируемой компетенции (или её части) и ее формулировка	Наименование оценочного средства	
<i>ОК-4: 3.1, 3.2, У.1</i> <i>ОК-4: В.1</i>	<i>тест, ИДЗ</i>	
Текущий контроль за выполнением лабораторных работ		

Модуль 2. Теория кодирования		
Цель – познакомить с основными методами кодирования		
<i>Содержание раздела</i>		
<i>План лекций</i>	<i>План лабораторных работ</i>	<i>План семинарских занятий</i>
<p>Лекция № 1. Алфавитное кодирование.</p> <p>1. Основные принципы.</p> <p>2. Минимизация длины кода сообщения.</p> <p>3. Неравенство МакМиллана.</p> <p>Лекция № 2. Сжатие информации.</p> <p>1. Основные принципы сжатия информации.</p> <p>2. Метод Лемпеля-Зива.</p> <p>Лекция № 3. Помехоустойчивое кодирование.</p> <p>1. Расстояние.</p> <p>2. Код Хемминга.</p> <p>Лекция № 4. Помехоустойчивое кодирование в системах сотовой связи.</p> <p>1. Расстояние.</p> <p>2. Код Хемминга.</p> <p>Методическое обеспечение лекций</p> <ul style="list-style-type: none"> • Презентации лекций хранятся на внутреннем портале университета; • Список литературы [1, 2, 4, 7, 12, 15, 16] 	<p>1. Кодирование информации методами Шеннона–Фено и Хаффмана (2 часа):</p> <ul style="list-style-type: none"> – метод Шеннона–Фено; – кодирование информации; – метод Хаффмана. <p>2. Сжатие данных по методу Лемпеля–Зива (2 часа):</p> <ul style="list-style-type: none"> – LZ-метод; – алгоритм LZ78. <p>3. Обнаружение одиночной ошибки методом Хемминга (2 часа):</p> <ul style="list-style-type: none"> – обнаружение одиночной ошибки (2 метода). <p>4. Защита реферата (помехоустойчивые коды) (2 часа)</p>	<p>Не предусмотрены учебным планом</p>

<i>Самостоятельная работа</i>		
Инвариантная часть	Вариативная часть	
<p>Темы, по которым осуществляется внеаудиторная работа студентов</p> <ul style="list-style-type: none"> • Алфавитное кодирование. • Помехоустойчивые и самокорректирующиеся коды. <p>Типы заданий, предлагаемых студентам</p> <ul style="list-style-type: none"> • Индивидуальное задание <p>Форма отчетности</p> <ul style="list-style-type: none"> • Сдача и защита индивидуального задания в течение 12 учебных дней со дня выдачи. <p>Текст заданий, предлагаемых студентам, хранится на внутреннем портале университета</p>	Сдача индивидуальных заданий досрочно +0,2	
Контроль (итоговый тест)		
<i>Паспорт оценочных средств по разделу</i>		
Код контролируемой компетенции (или её части) и ее формулировка	Наименование оценочного средства	
<i>ОК-4: 3.3; ОК-4: B.2</i>	<i>тест, 3 ИДЗ, реферат</i>	
Текущий контроль за выполнением лабораторных работ		
Модуль 3. Теория автоматов		
Цель – познакомить студентов с теорией автоматов, научить строить схемы автоматов		
<i>Содержание раздела</i>		
<i>План лекций</i>	<i>План лабораторных работ</i>	<i>План семинарских занятий</i>
<p>Лекция № 1. Автоматы.</p> <p>1. Логические элементы.</p> <p>2. Комбинационные схемы.</p>	<p>1. Комбинационные схемы (2 часа):</p> <p>– построение схем из логических элементов;</p>	<p>Не предусмотрены учебным планом</p>

<i>План лекций</i>	<i>План лабораторных работ</i>	<i>План семинарских занятий</i>
<p>3. Последовательные схемы. 4. Синтез автоматов. 5. Двоичный сумматор. Лекция № 2. Автоматы Мили и Мура. 1. Абстрактный автомат. 2. Структурный автомат. 3. Параллельное и последовательное соединение автоматов. Методическое обеспечение лекций</p> <ul style="list-style-type: none"> • Презентации лекций хранятся на внутреннем портале университета. • Список литературы [2, 3, 4, 7, 9, 10,16] 	<p>– упрощение логических уравнений; – построение субтрактора. 2. Регулярные выражения (2 часа): – запись регулярных событий; – применение методов абстрактного синтеза; – работа с программой «Регулярные выражения»</p>	<p>Не предусмотрены учебным планом</p>
<i>Самостоятельная работа</i>		
Инвариантная часть		Вариативная часть
<p>Темы, по которым осуществляется внеаудиторная работа студентов</p> <ul style="list-style-type: none"> • Задачи диагностики конечных автоматов. <p>Типы заданий, предлагаемых студентам</p> <ul style="list-style-type: none"> • Индивидуальное задание <p>Форма отчетности</p> <ul style="list-style-type: none"> • Сдача и защита индивидуального задания в течение 12 учебных дней со дня выдачи. 		<p>Сдача индивидуальных заданий досрочно +0,2</p>

Контроль (итоговый тест)		
<i>Паспорт оценочных средств по разделу</i>		
Код контролируемой компетенции (или её части) и ее формулировка	Наименование оценочного средства	
СК-1: 3.4, У.3; СК-1: В.3	<i>тест, собеседование ИДЗ, синквейн</i>	
Текущий контроль за выполнением лабораторных работ		
Модуль 4. Грамматики		
Цель – научить строить дерево разбора, выполнять грамматический разбор		
<i>Содержание раздела</i>		
<i>План лекций</i>	<i>План лабораторных работ</i>	<i>План семинарских занятий</i>
Лекция № 1. Грамматики. 1. Основные определения. 2. Классификация грамматик. 3. Локализованные структуры. 4. Канонические формы. 5. Деревья. Двоичные деревья трансляции. Общий алгоритм. Лекция № 2. Иерархия Хомского. 1. Классификация. 2. Конечный автомат. 3. Регулярные грамматики	1. Диаграммы Вирта (2 часа): – выделение синтаксических классов для описания языка Паскаль; – описание синтаксического класса «оператор». 2. Формальные грамматики. Разбор снизу вверх (алгоритм с возвратами) (2 часа): – построение дерева разбора; – алгоритм с возвратами	Не предусмотрены учебным планом

<i>План лекций</i>	<i>План лабораторных работ</i>	<i>План семинарских занятий</i>
<p>Лекция № 3. Регулярные выражения.</p> <p>1. Регулярные события.</p> <p>2. Методы абстрактного синтеза.</p> <p>3. Структурная теория автоматов.</p> <p>Методическое обеспечение лекций</p> <ul style="list-style-type: none"> • Презентации лекций хранятся на внутреннем портале университета. • Список литературы [3, 5, 8, 15, 17] 		<p>Не предусмотрены учебным планом</p>
<i>Самостоятельная работа</i>		
Инвариантная часть		Вариативная часть
<p>Темы, по которым осуществляется внеаудиторная работа студентов</p> <p>Типы заданий, предлагаемых студентам</p> <ul style="list-style-type: none"> • Индивидуальное задание <p>Текст заданий, предлагаемых студентам, хранится на внутреннем портале университета</p>		<p>Сдача индивидуальных заданий досрочно +0,2</p>
<i>Паспорт оценочных средств по разделу</i>		
Код контролируемой компетенции (или её части) и ее формулировка	Наименование оценочного средства	
СК-1: 3.5, У.3; СК-1: В.4, В.5	тест 2 ИДЗ, интеллект-карта	
Текущий контроль за выполнением лабораторных работ		

Модуль 5. Распознавание образов. Математическая кибернетика		
Цель – познакомить с теорией распознавания образов, рассмотреть основные понятия математической кибернетики		
<i>Содержание раздела</i>		
<i>План лекций</i>	<i>План лабораторных работ</i>	<i>План семинарских занятий</i>
<p>Лекция № 1. Распознавание образов и анализ сцен.</p> <ol style="list-style-type: none"> 1. Понятие образа. 2. Проблема обучения распознаванию образов (ОРО). 3. Геометрический и структурный подходы. 4. Обучение и самообучение <p>Лекция № 2. Математическая кибернетика.</p> <ol style="list-style-type: none"> 1. Основы теории управления. 2. Управление с обратной связью. 3. Теория принятия решений. <p>Методическое обеспечение лекции</p> <ul style="list-style-type: none"> • Презентации лекций хранятся на внутреннем портале университета. • Список литературы [1, 2, 9, 13, 14, 15, 16] 	<p>1. Распознавание образов (2 часа): методы распознавания образов; работа с программой <i>Prog</i></p>	<p>Не предусмотрены учебным планом</p>

<i>Самостоятельная работа</i>	
Инвариантная часть	Вариативная часть
<p>Темы, по которым осуществляется внеаудиторная работа студентов</p> <ul style="list-style-type: none"> • Синтез модели эвристического алгоритма распознавания. • Математическая постановка задачи распознавания изображений. <p>Текст заданий, предлагаемых студентам, хранится на внутреннем портале университета</p>	Сдача индивидуальных заданий досрочно +0,2
Контроль (итоговый тест)	
<i>Паспорт оценочных средств по разделу</i>	
Код контролируемой компетенции (или её части) и ее формулировка	Наименование оценочного средства
<i>СК-1: 3.6;</i> <i>СК-1: У.4</i>	<i>тест, доклад</i>
Текущий контроль за выполнением лабораторных работ	

Приложение 2

Содержание самостоятельной работы

Раздел	Тема для самостоятельного изучения	Задание для самостоятельного выполнения студентом	Кол-во часов	Методическое обеспечение	Форма отчетности
1.	<p>Понятие об информации. Количество информации. Содержательность и полезность информации. Передача дискретных сообщений по каналу без шумов и с шумами</p>	<p>Подготовка к тесту. Выполнение индивидуальных заданий по теме Подходы к измерению количества информации</p>	8	<p>Челябинский государственный педагогический университет (внутренний портал) > Учебно-методические материалы > inform > Bakalavriat > Inf> 4k > TOI</p>	Тест, ИДЗ
2.	<p>Теория кодирования</p>	<p>Выполнение индивидуальных заданий по теме: 1. Методы Шеннона–Фено и Хаффмена; 2. Сжатие данных. Метод Лемпеля-Зива; 3. Коды Хемминга.</p> <p>Подготовка реферата по помехоустойчивым кодам.</p>	8	<p>Челябинский государственный педагогический университет (внутренний портал) > Учебно-методические материалы > inform > Bakalavriat > Inf> 4k > TOI</p>	<p>Тест, 3 ИДЗ, реферат, выступление с сообщением по теме «Помехоустойчивое кодирование в системах сотовой связи»</p>

Раздел	Тема для самостоятельного изучения	Задание для самостоятельного выполнения студентом	Кол-во часов	Методическое обеспечение	Форма отчетности
3.	Теория автоматов	Подготовка к тесту. Выполнение индивидуальных заданий по теме: Схемы автоматов	12	Челябинский государственный педагогический университет (внутренний портал) > Учебно-методические материалы > inform > Bakalavriat > Inf> 4k > TOI	Тест, ИДЗ, синквейн
4.	Граматики	Подготовка к тесту. Выполнение индивидуальных заданий по теме: 1. Диаграммы Вирта; 2. Грамматический разбор сверху-вниз	8	Челябинский государственный педагогический университет (внутренний портал) > Учебно-методические материалы > inform > Bakalavriat > Inf> 4k > TOI	Тест, 2 ИДЗ, интеллект-карта
5.	Распознавание образов. Математическая кибернетика	Подготовка к тесту	9	Челябинский государственный педагогический университет (внутренний портал) > Учебно-методические материалы > inform > Bakalavriat > Inf> 4k > TOI	Тест, доклад

Приложение 3

Рейтинг

Компетенции		ОК-4 3.1, 3.2		ОК-4 В.1		ОК-4 У.1		ОК-4 У.1		ОК-4		ОК-4 3.3		ОК-4 В.2		ОК-4 В.2		ОК-4 В.2		ОК-4	
№	Фамилия, Имя	Тест 1		Анализ текстов		Измерение кол-ва л/р		Статистич. тесты л/р		Модуль1	Тест 2		Шеннона-Фено		Хаффмена		Лемпел-Зива		Модуль2		
		вес	коэфф.	вес	коэфф.	вес	коэфф.	вес	коэфф.		вес	коэфф.	вес	коэфф.	вес	коэфф.	вес	коэфф.			
1	Кузванова Настя	60	0,69	14	1,2	13	1	13	1	84,2	60	0,63	13	1	14	1	13	1,2	80,4		
2	Турбабин Федор	60	0,81	14	1	13	1	13	1	88,6	60	0,63	13	1,2	14	1,2	13	1,2	85,8		
3	Истомин Стас	60	0,78	14	1	13	1	13	1	86,8	60	0,88	13	1	14	1	13	1,2	95,4		
4	Бердышев Никита	60	0,63	14	1	13	1	13	1	77,8	60	0,69	13	1	14	1	13	1,2	84		
5	Сагитова Динара	60	0,56	14	0,5	13	1	13	1	66,6	60	0,63	13	0,5	14	0,5	13	1	64,3		
6	Султанова Гульсия	60	0,63	14	1	13	1	13	1	77,8	60	0,69	13	1	14	1	13	1,2	84		
7	Ханова Гульфия	60	0,67	14	1	13	1	13	1	80,2	60	0,88	13	1	14	1	13	1,2	95,4		
8	Жукова Анжелика	60	0,67	14	1	13	1	13	1	80,2	60	0,38	13	1	14	1	13	1,2	65,4		
9	Семенов Максим	60	0,75	14	1	13	1	13	1	85	60	0,5	13	0,5	14	0,5	13	1,2	59,1		
10	Удалов Виталий	60	0,81	14	1,2	13	1	13	1	91,4	60	0,88	13	1	14	1	13	1,2	95,4		
11	Бортник Настя	60	0,63	14	1	13	1	13	1	77,8	60	0,38	13	1	14	1	13	1,2	65,4		
12	Семенов Федор	60	0,67	14	1	13	1	13	1	80,2	60	0,88	13	1,2	14	1,2	13	1,2	100,8		
13	Фомина Екатерина	60	0,63	14	1	13	1	13	1	77,8	60	0,69	13	1	14	1	13	1,2	84		

Продолжение

СК-1 3.4		ОК-4 У.3		СК-1 В.3		ОК-4 У.3		ОК-4, СК-1		СК-1 3.5		СК-1 У.3		СК-1 В.4		СК-1 В.5		СК-1	
Тест 3		Помехоуст-вое кодирование		Автоматы		Код Хемминга л/р		Модуль3	Тест 4		Регулярные выражения л/р		Диаграммы Вирта		Грамматик		Модуль4		
вес	коэфф.	вес	коэфф.	вес	коэфф.	вес	коэфф.		вес	коэфф.	вес	коэфф.	вес	коэфф.	вес	коэфф.			
60	0,56	13	1,4	14	1,2	13	1,2	84,2	60	0,82	10	1	10	1,4	20	1	93,2		
60	0,63	13	1,2	14	1,2	13	1,2	85,8	60	0,88	10	1	10	1	20	1,2	96,8		
60	0,75	13	1,2	14	1,2	13	1,2	93	60	0,75	10	1	10	1	20	1	85		
60	0,65	13	1,2	14	1,2	13	1,2	87	60	0,81	10	1	10	1	20	1	88,6		
60	0,83	13	1	14	1	13	1,2	92,4	60	0,69	10	1	10	1	20	1	81,4		
60	0,69	13	1	14	1	13	1,2	84	60	0,56	10	1	10	1	20	0,8	69,6		
60	0,88	13	1	14	1,2	13	1,2	98,2	60	0,5	10	1	10	1	20	1	70		
60	0,88	13	0,8	14	1	13	1,2	92,8	60	0,56	10	1	10	1	20	1	73,6		
60	0,75	13	1	14	1	13	1,2	87,6	60	0,63	10	1	10	1	20	1	77,8		
60	0,93	13	1,4	14	1,2	13	1,2	106,4	60	0,89	10	1	10	1,2	20	1,2	99,4		
60	0,67	13	1	14	1,2	13	1,2	85,6	60	0,56	10	1	10	1	20	1	73,6		
60	0,88	13	1	14	1	13	1	92,8	60	0,81	10	1	10	1	20	1	88,6		
60	0,75	13	1	14	1	13	1	85	60	0,88	10	1	10	1	20	1	92,8		

Окончание

СК-1	У.4	СК-1	З.6	СК-1					Компетенции	
Распозн. образов л/р		Тест 5		Модуль	ОК-4	СК-1	Итого	Оценка	№	Фамилия, Имя
вес	коэфф.	вес	коэфф.							
40	1	60	0,8	88	82,93	88,47	85,70	хорошо	1	Кузеванова Настя
40	1	60	0,94	96,4	86,73	93,00	89,87	отлично	2	Турбабин Федор
40	1	60	0,75	85	91,73	87,67	89,70	отлично	3	Истомин Стас
40	1	60	0,63	77,8	82,93	84,47	83,70	хорошо	4	Бердышев Никита
40	1	60	0,56	73,6	74,43	82,47	78,45	удовл.	5	Сагитова Динара
40	1	60	0,56	73,6	81,93	75,73	78,83	удовл.	6	Султанова Гульсия
40	1	60	0,48	68,8	91,27	79,00	85,13	хорошо	7	Ханова Гульфия
40	1	60	0,75	85	79,47	83,80	81,63	хорошо	8	Жукова Анжелика
40	1	60	0,48	68,8	77,23	78,07	77,65	удовл.	9	Семенов Максим
40	1	60	0,94	96,4	97,73	100,73	99,23	отлично	10	Удалов Виталий
40	1	60	0,63	77,8	76,27	79,00	77,63	удовл.	11	Бортник Настя
40	1	60	0,8	88	91,27	89,80	90,53	отлично	12	Семенов Федор
40	1	60	0,63	77,8	82,27	85,20	83,73	хорошо	13	Фомина Екатерина

Анализ текстов

Анализ текстовой информации в настоящее время является актуальной задачей. Применение статистики для анализа текстов – традиционная задача¹.

Рассмотрим некоторые интересные факты относительно частоты встречаемости букв и их сочетаний в разных языках (подробнее см. например, недавно вышедшую интересную книгу Анализ текстов²).

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 $T = t_1t_2\dots t_l$, составленных из букв алфавита $\{a_1, a_2, \dots, a_n\}$. При этом просматриваются подряд идущие m -граммы текста:

$$t_1t_2\dots t_m, t_2t_3\dots t_{m+1}, \dots, t_{i-m+1}t_{i-m+2}\dots t_i.$$

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

¹ Анализ текстов. Режим доступа. 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. – М.: Гелиос АРВ, 2002.

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

для данной m -граммы мало отличаются друг от друга.

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

Ниже приводится таблица частот букв (в процентах) ряда европейских языков. Данные заимствованы из книги Elements de cryptographie³.

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

³ Baudouin C. Elements de cryptographie / Ed. Pedone A. Paris, 1939.

Буква алфавита	Французский язык	Немецкий язык	Английский язык	Испанский язык	Итальянский язык
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

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.

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*.

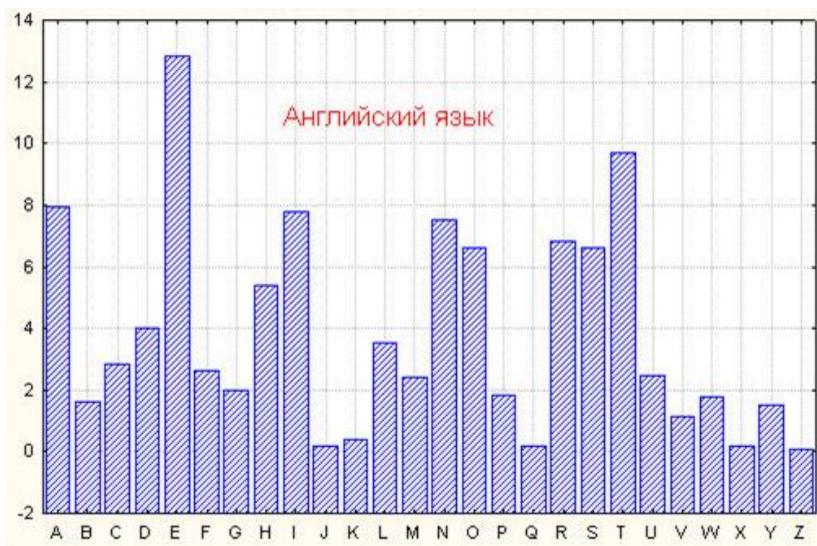


Рис. 1. Диаграмма частоты использования букв алфавита в английском языке

Для русского языка частоты (в порядке убывания) знаков алфавита, в котором отождествлены *Е* с *Ё*, *Б* с *Ь*, а также имеется знак пробела (–) между словами, приведены в следующей таблице (см. в книге Вероятность и информация⁴).

–	О	Е, Ё	А	И	Т	Н	С
0,175	0,090	0,072	0,062	0,062	0,053	0,053	0,045
Р	В	Л	К	М	Д	П	У
0,040	0,038	0,035	0,028	0,026	0,025	0,023	0,021
Я	Ы	З	Ь, Ь	Б	Г	Ч	Й
0,018	0,016	0,016	0,014	0,014	0,013	0,012	0,010
Х	Ж	Ю	Ш	Ц	Щ	Э	Ф
0,009	0,007	0,006	0,006	0,004	0,003	0,003	0,002

⁴ Яглом А.М. Вероятность и информация / А.М. Яглом, И.М. Яглом. М.: Наука, 1973.

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

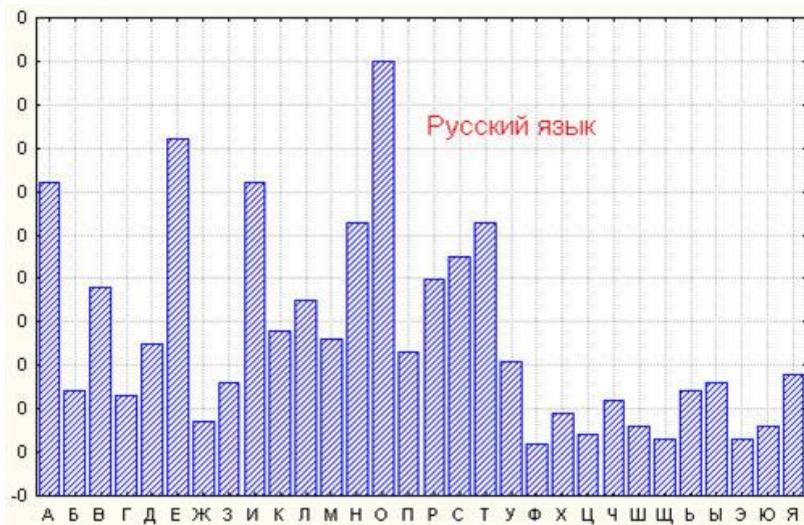


Рис. 2. Диаграмма частот использования букв алфавита для русского языка

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).

Французский язык	Е, S, A, N, T, I, R, U, L, O	79,9%
Немецкий язык	Е, N, I, S, T, A, H, D, U	77,2%
Английский язык	Е, T, A, I, N, R, O, S, H, D	75,3%
Испанский язык	Е, А, О, S, I, R, N, L, D, С	78,3%
Итальянский язык	І, Е, А, О, N, Т, R, L, S, Т	79,9%

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

Приведем таблицы частот биграмм для русского и английского языков (таблицы заимствованы из книги *Military cryptanalysis*⁵). Для удобства они разбиты на четыре части по следующей схеме:

Часть1	Часть2
Часть3	Часть4

Часть1																
	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	

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

Часть2															
	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

Часть3																
	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

Часть4															
	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

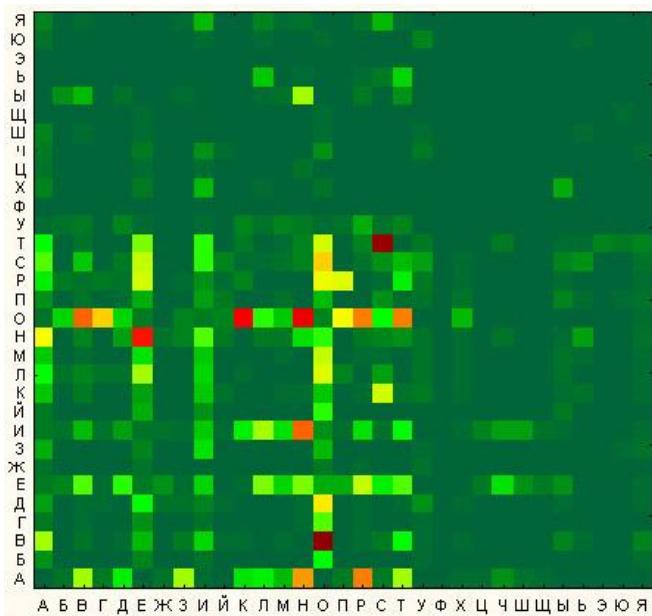


Рис. 3. График частот использования пар букв русского алфавита

Хорошие таблицы k -грамм легко получить, используя тексты электронных версий многих книг, содержащихся на CD-дисках.

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.

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

Окончание таблицы

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

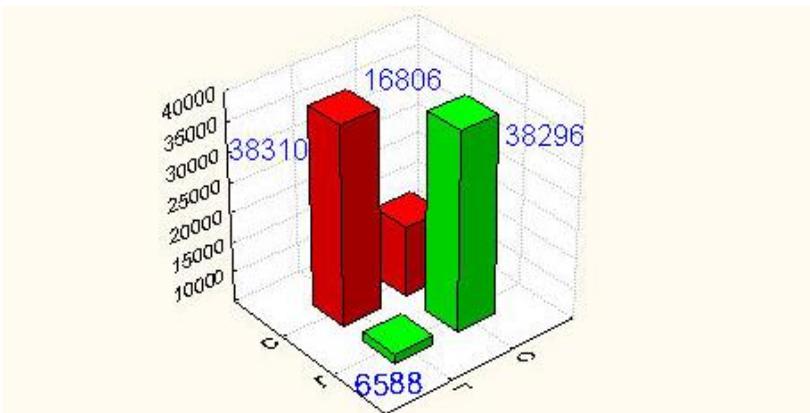


Рис. 4. Частота чередования гласных и согласных букв

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

$$p(a_{i1}) \neq p(a_{i1}/a_{i2}), p(a_{i1}/a_{i2}) \neq p(a_{i1}/a_{i2}/a_{i3}), \dots$$

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 (z, z), гласная-согласная (z, c), согласная-гласная (c, z), согласная-согласная (c, c) в русском тексте длиной в 10^5 знаков. Результаты подсчета отражены в следующей таблице:

	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:

$$p(z/c) \approx 0.663, p(c/z) \approx 0.872, p(z) \approx 0.432, p(c) \approx 0.568.$$

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.

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%

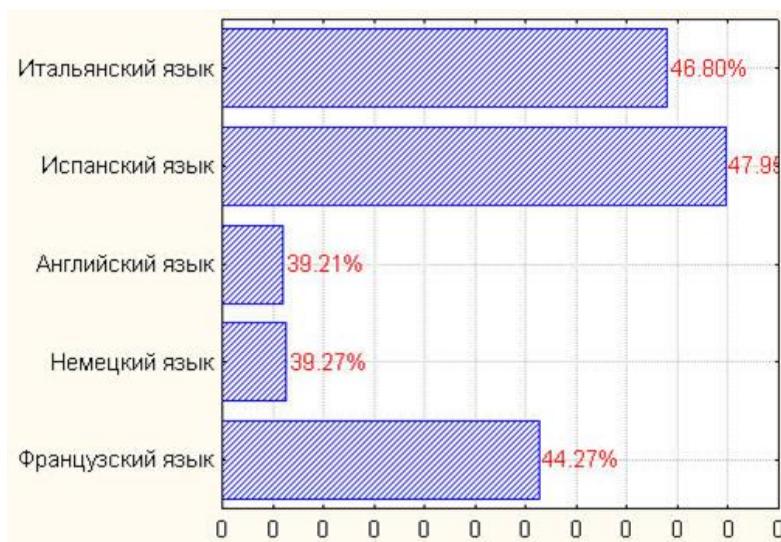


Рис. 4. Доля гласных букв в алфавитах

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 ленточного принтера – это замечательное воплощение идеи ускорения набора текста, связанное с оптимизацией расположения букв алфавита относительно друг друга в зависимости от частоты их применения.

Пример текста для анализа

Information Theory⁶

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. $\text{Prob}(S = i) = p_i$. 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 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

⁶ Information Theory. <http://bactra.org/notebooks/information-theory.html>.

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 preferred 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

ТЕОРЕТИЧЕСКИЕ ОСНОВЫ ИНФОРМАТИКИ
Практикум

ISBN 978-5-906777-56-0

Работа рекомендована РИСом университета
Протокол № 8, пункт 20, 2015 г.

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

Редактор О.В. Максимова
Компьютерная верстка Т.Н. Никитенко

Формат 60×84/16. Объем 7,5 уч.-изд. л.
Бумага типографская. Подписано в печать 5.11.2015 г.
Тираж 100 экз. Заказ № 768

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