ГОСУДАРСТВЕННОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«БЕЛОРУССКО-РОССИЙСКИЙ УНИВЕРСИТЕТ»

Кафедра «Автоматизированные системы управления»

ТЕОРИЯ ИНФОРМАЦИИ

Методические рекомендации к лабораторным работам для студентов направлений подготовки 09.03.01 «Информатика и вычислительная техника» и 09.03.04 «Программная инженерия»

Могилев 2018

УДК 004. 3
ББК 32.973
Т 33

Рекомендовано к изданию
учебно-методическим отделом
Белорусско-Российского университета

Одобрено кафедрой «Автоматизированные системы управления» «04» сентября 2018 г., протокол № 2

Составитель ст. преподаватель А. С. Сидоренко

Рецензент доц. И. В. Лесковец

Даны методические рекомендации по выполнению лабораторных работ по дисциплине «Теория информации», а также приведены задания к ним и список литературы для подготовки.

Учебно-методическое издание

ТЕОРИЯ ИНФОРМАЦИИ

Ответственный за выпуск	А. И. Якимов
Технический редактор	С. Н. Красовская
Компьютерная верстка	Н. П. Полевничая

Подписано в печать . Формат 60×84/16. Бумага офсетная. Гарнитура Таймс.
Печать трафаретная. Усл. печ. л. . Уч.-изд. л. . Тираж 16 экз. Заказ №

Издатель и полиграфическое исполнение:
Государственное учреждение высшего профессионального образования
«Белорусско-Российский университет».
Свидетельство о государственной регистрации издателя,
изготовителя, распространителя печатных изданий
№ 1/156 от 24.01.2014.
Пр. Мира, 43, 212000, Могилев.

Содержание

Введение

Целью преподавания дисциплины «Теория информации» является формирование у студентов базовых знаний по теории информации и об основных её проблемах, возникающих при получении, обработке, передаче и использовании информации в системах различного назначения и практической деятельности человека, об основных методах оценки количества информации в непрерывных и дискретных сообщениях, об основных методах обеспечения верности и эффективности передачи информации в условиях помех и без помех по предоставленным каналам связи.

Методические рекомендации имеют целью помочь студентам в подготовке и выполнению лабораторных работ по дисциплине.

Даны методические рекомендации по выполнению лабораторных работ по дисциплине «Теория информации», а также приведены задания к ним и список литературы для подготовки.

Лабораторная работа № 1. Энтропия и ее свойства

Цель работы: изучение понятия энтропии как среднестатистической меры неопределенности знаний получателя информации.

Порядок выполнения работы

1 Изучить основные теоретические положения, сделав необходимые выписки в конспект.
2 Получить задание у преподавателя, выполнить типовые задания.
3 Сделать выводы по результатам исследований.
4 Оформить отчет.

Требования к отчету

1 Цель работы.
2 Постановка задачи.
3 Результаты исследования.
4 Выводы.

Основные теоретические положения

Количество информации, содержащееся в одном элементарном сообщении x_i, не полностью характеризует источник. Источник дискретных сообщений может быть охарактеризован средним количеством информации, приходящимся на одно элементарное сообщение, носящим название «энтропия источника»:

$$H(X) = -\sum_{i=1}^{k} p_i \log_2 p_i, \qquad i = 1...k, \tag{1}$$

где k – объём алфавита сообщений.

Таким образом, энтропия – это среднестатистическая мера неопределенности знаний получателя информации относительно состояния наблюдаемого объекта.

В выражении (1) статистическое усреднение (т. е. определение математического ожидания дискретной случайной величины $I(Xi)$) выполняется по всему ансамблю сообщений источника. При этом необходимо учитывать все вероятностные связи между сообщениями. Чем выше энтропия источника, тем большее количество информации в среднем закладывается в каждое сообщение, тем труднее запомнить (записать) или передать такое сообщение по каналу связи. Таким образом, суть энтропии Шеннона заключается в следующем: энтропия дискретной случайной величины – это минимум среднего количества битов, которое нужно передавать по каналу связи о текущем значении данной случайной величины.

Необходимые затраты энергии на передачу сообщения пропорциональны энтропии (среднему количеству информации на сообщение). Отсюда следует,

что количество информации в последовательности из N сообщений определяется количеством этих сообщений и энтропией источника, т. е.

$$I(N) = NH(X).$$

Энтропия как количественная мера информационности источника обладает следующими свойствами:

– энтропия равна нулю, если хотя бы одно из сообщений достоверно (т. е. имеет вероятность $p_i = 1$);

– величина энтропии всегда больше или равна нулю, действительна и ограничена;

– энтропия источника с двумя альтернативными событиями может изменяться от 0 до 1;

– энтропия – величина аддитивная: энтропия источника, сообщения которого состоят из сообщений нескольких статистически независимых источников, равна сумме энтропий этих источников;

– энтропия будет максимальной, если все сообщения равновероятны:

$$H_{\max}(x) = -\frac{1}{k}\sum_{i=1}^{k}\log_2\frac{1}{k} = \log_2 k\,. \qquad (2)$$

При неравновероятных сообщениях x_i энтропия уменьшается. В связи с этим вводят такую меру источника, как статистическая избыточность алфавита источника

$$\rho_x = 1 - \frac{H(x)}{H(X)_{\max}} = 1 - \frac{H(X)}{\log_2 k}\,, \qquad (3)$$

где $H(X)$ – энтропия реального источника;

$H(X)_{\max}$ – максимально достижимая энтропия источника, $H(X)_{\max} = \log_2$.

Определяемая по формуле (3) избыточность источника информации говорит об информационном резерве сообщений, элементы которых неравновероятны.

Существует также понятие семантической избыточности, которое следует из того, что любую мысль, которая содержится в сообщении из предложений человеческого языка, можно сформулировать короче. Считается, что если какое-либо сообщение можно сократить без потери его смыслового содержания, то оно имеет семантическую избыточность.

Рассмотрим дискретные случайные величины (д. с. в.) X и Y, заданные законами распределения $P(X = X_i) = p_i$, $P(Y = Y_j) = q_j$ и совместным распределением $P(X = X_i, Y = Y_j) = p_{ij}$. Тогда количество информации, содержащееся в д. с. в. X относительно д. с. в. Y, определяется по формуле

$$(X,Y)=\sum_{ij}\rho_{ij}\log_2\frac{\rho_{ij}}{\rho_i q_j}\ . \qquad (4)$$

Для непрерывных случайных величин (сл. в.) X и Y, заданных плотностями распределения вероятностей $r_X(t_1)$, $r_Y(t_2)$ и $r_{XY}(t_1,t_2)$, аналогичная формула имеет вид:

$$I(X,Y)=\iint_{R^2}\rho_{XY}(t_1,t_2)\log_2\frac{\rho_{XY}(t_1,t_2)}{\rho_X(t_1)\rho_Y(t_2)}dt_1dt_2\,.$$

Очевидно, что

$$P\left(X=X_i,X=X_j\right)=\begin{cases}0, & \text{при } i\neq j;\\ P\left(X=X_i\right), & \text{при } i=j,\end{cases}$$

следовательно

$$(X,X)=\sum p_i\log_2\frac{p_i}{p_i p_i}=-\sum p_i\log_2 p_i\,,$$

т. е. приходим к выражению (1.1) для расчета энтропии $H(X)$.

Задания для самостоятельной работы

1 Имеются два ящика, в каждом из которых по 12 шаров. В первом – три белых, три черных и шесть красных; во втором – каждого цвета по четыре. Опыты состоят в вытаскивании по одному шару из каждого ящика. Что можно сказать относительно неопределенностей исходов этих опытов?

2 В ящике имеются два белых шара и четыре черных. Из ящика извлекают последовательно два шара без возврата. Найти энтропию, связанную с первым и вторым извлечениями, а также энтропию обоих извлечений.

3 Имеется три тела с одинаковыми внешними размерами, но с разными массами x_1, x_2 *и* x_3. Необходимо определить энтропию, связанную с нахождением наиболее тяжелого из них, если сравнивать веса тел можно только попарно.

4 Какое количество информации требуется, чтобы узнать исход броска монеты?

5 Случайным образом вынимается карта из колоды в 32 карты. Какое количество информации требуется, чтобы угадать, что это за карта? Как построить угадывание?

6 В некоторой местности имеются две близкорасположенные деревни: А и В. Известно, что жители А всегда говорят правду, а жители В – всегда лгут. Известно также, что жители обеих деревень любят ходить друг к другу в гости, поэтому в каждой из деревень можно встретить жителя соседней деревни. Путешественник, сбившись ночью с пути оказался в одной из двух деревень и, заговорив с первым встречным, захотел выяснить, в какой деревне он находится и

откуда его собеседник. Какое минимальное количество вопросов с бинарными ответами требуется задать путешественнику?

7 В Петрозаводске 280000 жителей. Какое минимальное количество вопросов, требующих ответа «да» или «нет», необходимо, чтобы однозначно найти одного жителя.

8 В лотерее N билетов, из них K выигрышных. Студент купил M билетов и после розыгрыша сообщил Вам, что выиграл (но, возможно, и не на один билет). Какое количество информации Вы получили?

Вопросы для контроля

1 Какие существуют виды информации?
2 Как перевести непрерывную информацию в дискретный (цифровой) вид?
3 Что такое частота дискретизации непрерывной информации?
4 Как формулируется теорема дискретизации?
5 Что такое информация, кодирование, канал связи, шум?
6 В чем заключаются основные положения вероятностного подхода Шеннона к определению количества информации?
7 Как определяется количество информации, содержащееся в одном сообщении дискретного источника?
8 Как определяется количество информации на одно сообщение источника взаимозависимых сообщений?
9 Что такое энтропия источника? Какие ее свойства?
10 При каких условиях энтропия источника максимальна?
11 Как определяется количество информации? Какие свойства количества информации?
12 Чем обусловлена статистическая избыточность источника информации?

Лабораторная работа № 2. Количество информации

Цель работы: изучение количественных характеристик информации как объекта передачи данных.

Порядок выполнения работы

1 Изучить основные теоретические положения, сделав необходимые выписки в конспект.
2 Получить задание у преподавателя, выполнить типовые задания.
3 Сделать выводы по результатам исследований.
4 Оформить отчет.

Требования к отчету

1 Цель работы.
2 Постановка задачи.

3 Результаты исследования.
4 Выводы.

Основные теоретические положения

Информация – это сведения, являющиеся объектом передачи, распределения, преобразования, хранения или непосредственного использования.

Сообщение является формой представления информации.

Одно и то же сведение может быть представлено в различной форме. Например, сведение о моменте начала наступления может быть передано по телефону или телеграфом или тремя зелеными ракетами. В первом случае мы имеем дело с информацией, представленной в непрерывном виде (непрерывное сообщение). Будем считать, что это сообщение вырабатывается источником непрерывных сообщений. Во втором и третьем случаях – с информацией, представленной в дискретном виде (дискретное сообщение). Это сообщение вырабатывается источником дискретных сообщений.

Основное отличие дискретного и непрерывного источников состоит в следующем. Множество всех различных сообщений, вырабатываемых дискретным источником всегда конечно. Поэтому на конечном отрезке времени количество символов дискретного источника также является конечным. В то же время число возможных различных значений звукового давления (или напряжения в телефонной линии), измеренное при разговоре, даже на конечном отрезке времени, будет бесконечным.

В нашем курсе мы будем рассматривать вопросы передачи именно дискретных сообщений. При этом в случае телефонной связи под сообщением будем понимать некоторую последовательность отсчетов квантованного аналогового сигнала, передаваемую в канале связи в виде последовательности кодовых комбинаций.

Информация, содержащаяся в сообщении, передается от источника сообщений к получателю по каналу передачи дискретных сообщений (ПДС) (рисунок 1).

Рисунок 1 – Тракт передачи дискретных сообщений

Сообщение поступает от источника дискретных сообщений, который характеризуется алфавитом передаваемых сообщений $A = \{a_1, a_2, ..., a_n\}$.

Алфавит – есть совокупность всех возможных (различных) сообщений (знаков) данного источника.

Объем алфавита – число различных символов алфавита K.

Каждое сообщение алфавита появляется с некоторой вероятностью.

Вероятность выдачи символа (сообщения) a_i – $p(a_i)$.

Количество информации в сообщении (символе) определяется вероятностью его появления. Чем меньше вероятность появления того или иного сообщения, тем большее количество информации мы извлекаем при его получении. В 1928 г. Хартли предложил определять количество информации, которое приходится на одно сообщение a_i, выражением

$$I(a_i) = \log_2 \frac{1}{p(a_i)} = -\log_2 p(a_i).$$

Один бит – это количество информации, которое переносит один символ источника дискретных сообщений в том случае, когда алфавит источника состоит из двух равновероятных символов.

Среднее количество информации, выдаваемое источником в единицу времени, называют производительностью источника

$$H^*(A) = \frac{H(A)}{t_{cp}},$$

где t_{cp} – среднее время, отводимое на передачу одного символа (сообщения).

Среднее время может быть определено выражением

$$t_{cp} = \sum_{i=1}^{K} p(a_i) t_i .$$

Сигналы – форма сообщения для передачи по каналу связи.

Любая система связи обеспечивает передачу именно сигналов, а не сообщений. Поэтому сообщение, поступающее от источника, предварительно должно быть преобразовано в сигнал определенной природы (электрический, оптический, …), который является его переносчиком в данной системе связи.

Виды сигналов. Различают четыре вида сигналов: непрерывный непрерывного времени, непрерывный дискретного времени, дискретный непрерывного времени и дискретный дискретного времени.

Непрерывные сигналы непрерывного времени называют сокращенно непрерывными (аналоговыми) сигналами. Они могут изменяться в произвольные моменты, принимая любые значения из непрерывного множества возможных значений. К таким сигналам относится и известная всем синусоида.

Вопросы для контроля

1 Дайте определение понятиям «Информация» и «сообщение».
2 Перечислите основные характеристики источника дискретных сообщений.
3 Чем определяется количество информации в дискретном сообщении?
4 Что такое энтропия источника и как она определяется?
5 Как определить производительность дискретного источника?
6 Дайте определение основным параметрам цифровых сигналов данных.

Лабораторная работа № 3. Простейшие алгоритмы сжатия информации

Цель работы: изучение простейших алгоритмов сжатия информации.

Порядок выполнения работы

1 Изучить основные теоретические положения, сделав необходимые выписки в конспект.
2 Получить задание у преподавателя, выполнить типовые задания.
3 Сделать выводы по результатам исследований.
4 Оформить отчет.

Требования к отчету

1 Цель работы.
2 Постановка задачи.
3 Результаты исследования.
4 Выводы.

Основные теоретические положения

Алгоритм RLE.

Данный алгоритм необычайно прост в реализации. Групповое кодирование – от английского Run Length Encoding (RLE) – один из самых старых и самых простых алгоритмов архивации графики. Изображение в нем (как и в нескольких алгоритмах, описанных ниже) вытягивается в цепочку байт по строкам растра. Само сжатие в RLE происходит за счет того, что в исходном изображении встречаются цепочки одинаковых байт. Замена их на пары <счетчик повторений, значение> уменьшает избыточность данных.

Алгоритм декомпрессии при этом выглядит так:

```
Initialization(...);
do {
   byte = ImageFile.ReadNextByte();
   if(является счетчиком(byte)) {
      counter = Low6bits(byte)+1;
      value = ImageFile.ReadNextByte();
      for(i=1 to counter)
         DecompressedFile.WriteByte(value)
      }
   else {
      DecompressedFile.WriteByte(byte)
} while(ImageFile.EOF());
```

В данном алгоритме признаком счетчика (counter) служат единицы в двух верхних битах считанного файла (рисунок 2).

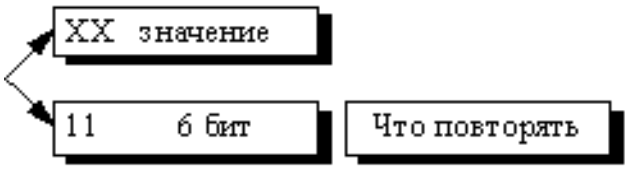

Рисунок 2 – Признаки счетчика

Соответственно оставшиеся 6 бит расходуются на счетчик, который может принимать значения от 1 до 64. Строку из 64 повторяющихся байтов превращают в два байта, т. е. сжимают в 32 раза.

Второй вариант алгоритма.

Второй вариант этого алгоритма имеет больший максимальный коэффициент архивации и меньше увеличивает в размерах исходный файл.

Алгоритм декомпрессии для него выглядит так:

```
Initialization(...);
do {
   byte = ImageFile.ReadNextByte();
   counter = Low7bits(byte)+1;
   if(если признак повтора(byte)) {
      value = ImageFile.ReadNextByte();
      for (i=1 to counter)
      CompressedFile.WriteByte(value)
   }
   else {
   for(i=1 to counter){
      value = ImageFile.ReadNextByte();
      CompressedFile.WriteByte(value)
   }
   CompressedFile.WriteByte(byte)
} while(ImageFile.EOF());
```

Признаком повтора в данном алгоритме является единица в старшем разряде соответствующего байта (рисунок 3).

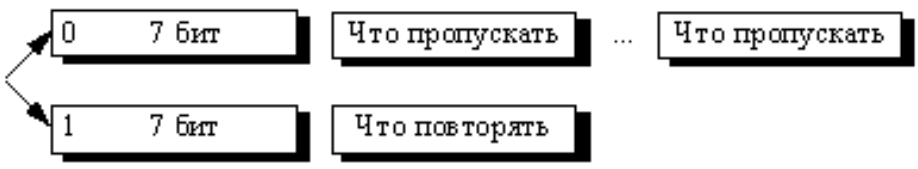

Рисунок 3 – Признак повтора

Как можно легко подсчитать, в лучшем случае этот алгоритм сжимает файл в 64 раза (а не в 32 раза, как в предыдущем варианте), в худшем увеличивает на 1/128. Средние показатели степени компрессии данного алгоритма находятся на уровне показателей первого варианта.

Алгоритм LZ.

Существует довольно большое семейство *LZ*-подобных алгоритмов, различающихся, например, методом поиска повторяющихся цепочек. Один из достаточно простых вариантов этого алгоритма, например, предполагает, что во входном потоке идет либо пара <счетчик, смещение относительно текущей позиции>, либо просто <счетчик> «пропускаемых» байт и сами значения байтов (как во втором варианте алгоритма *RLE*). При разархивации для пары <счетчик, смещение> копируются <счетчик> байт из выходного массива, полученного в результате разархивации, на <смещение> байт раньше, а <счетчик> (т. е. число, равное счетчику) значений «пропускаемых» байт просто копируются в выходной массив из входного потока. Данный алгоритм является несимметричным по времени, поскольку требует полного перебора буфера при поиске одинаковых подстрок. В результате сложно задать большой буфер из-за резкого возрастания времени компрессии. Однако потенциально построение алгоритма, в котором на <счетчик> и на <смещение> будет выделено по 2 байта (старший бит старшего байта счетчика – признак повтора строки / копирования потока), дает возможность сжимать все повторяющиеся подстроки размером до 32 Кб в буфере размером 64 Кб.

При этом получается увеличение размера файла в худшем случае на 32770/32768 (в двух байтах записано, что нужно переписать в выходной поток следующие 215 байт), что совсем неплохо. Максимальный коэффициент сжатия составит в пределе 8192 раза. В пределе, поскольку максимальное сжатие получают, превращая 32 Кб буфера в 4 байта, а буфер такого размера накапливается не сразу. Однако, минимальная подстрока, для которой становится выгодным проводить сжатие, должна состоять в общем случае минимум из 5 байт, что и определяет малую ценность данного алгоритма. К достоинствам *LZ* можно отнести чрезвычайную простоту алгоритма декомпрессии.

Классический алгоритм Хаффмана.

Один из классических алгоритмов, известных с 60-х гг. Использует только частоту появления одинаковых байт в изображении. Сопоставляет символам входного потока, которые встречаются большее число раз, цепочку бит меньшей длины. И, напротив, встречающимся редко – цепочку большей длины. Для сбора статистики требует двух проходов по изображению.

Задания для самостоятельной работы

1 Составьте алгоритм компрессии для первого варианта алгоритма *RLE*.
2 Составьте алгоритм компрессии для второго варианта алгоритма *RLE*.

3 Предложите другой вариант алгоритма *LZ*, в котором на пару <счетчик, смещение> будет выделено 3 байта, и подсчитайте основные характеристики своего алгоритма.

Вопросы для контроля

1 Предложите два-три примера «плохих» изображений для алгоритма *RLE*. Объясните, почему размер сжатого файла больше размера исходного файла

2 На какой класс изображений ориентирован алгоритм *RLE*?

3 Приведите пример «плохого» изображения для алгоритма Хаффмана.

4 Сравните алгоритмы сжатия изображений без потерь.

Лабораторная работа № 4. Арифметическое кодирование

Цель работы: изучение основных принципов арифметического кодирования.

Порядок выполнения работы

1 Изучить основные теоретические положения, сделав необходимые выписки в конспект.

2 Получить задание у преподавателя, выполнить типовые задания.

3 Сделать выводы по результатам исследований.

4 Оформить отчет.

Требования к отчету

1 Цель работы.

2 Постановка задачи.

3 Результаты исследования.

4 Выводы.

Основные теоретические положения

При арифметическом кодировании текст представляется вещественными числами в интервале от 0 до 1. По мере кодирования текста, отображающий его интервал уменьшается, а количество битов для его представления возрастает. Очередные символы текста сокращают величину интервала, исходя из значений их вероятностей, определяемых моделью. Более вероятные символы делают это в меньшей степени, чем менее вероятные, и, следовательно, добавляют меньше битов к результату.

Перед началом работы соответствующий тексту интервал есть [0; 1). При обработке очередного символа его ширина сужается за счет выделения этому символу части интервала. Например, применим к тексту «eaii!» алфавита {a, e, i, o, u, !} модель с постоянными вероятностями, заданными в таблице 1.

Таблица 1 – Пример постоянной модели для алфавита { a,e,i,o,u,! }

Символ	Вероятность	Интервал
a	.2	[0.0;0.2)
e	.3	[0.2;0.5)
i	.1	[0.5;0.6)
o	.2	[0.6;0.8)
u	.1	[0.8;0.9)
!	.1	[0.9;1.0)

И кодировщику и декодировщику известно, что в самом начале интервал – есть [0; 1). После просмотра первого символа «e», кодировщик сужает интервал до [0.2; 0.5), который модель выделяет этому символу. Второй символ «a» сузит этот новый интервал до первой его пятой части, поскольку для «a» выделен фиксированный интервал [0.0; 0.2). В результате получим рабочий интервал [0.2; 0.26), т. к. предыдущий интервал имел ширину в 0.3 единицы и одна пятая от него есть 0.06. Следующему символу «i» соответствует фиксированный интервал [0.5; 0.6), что применительно к рабочему интервалу [0.2; 0.26) суживает его до интервала [0.23, 0.236). Продолжая в том же духе, имеем листинг результата программы:

```
В начале              [0.0;     1.0   )
  После просмотра "e"    [0.2;     0.5   )
    -"-"-"-     "a"    [0.2;     0.26  )
    -"-"-"-     "i"    [0.23;    0.236 )
    -"-"-"-     "i"    [0.233;   0.2336)
    -"-"-"-     "!"    [0.23354; 0.2336)
```

Предположим, что все что декодировщик знает о тексте, это конечный интервал [0,23354; 0,2336). Он сразу же понимает, что первый закодированный символ есть «e», т. к. итоговый интервал целиком лежит в интервале, выделенном моделью этому символу согласно таблице 1. Теперь повторим действия кодировщика и получиим результата программы:

```
Сначала               [0.0; 1.0)
После просмотра "e"    [0.2; 0.5)
```

Отсюда ясно, что второй символ – это «a», поскольку это приведет к интервалу [0.2; 0.26), который полностью вмещает итоговый интервал [0,23354; 0,2336). Продолжая работать таким же образом, декодировщик извлечет весь текст.

Задание для самостоятельной работы

С помощью арифметического кодирования требуется закодировать сообщение «информационный», а затем представить сообщение в двоичном виде.

Список литературы

1 **Баскаков, С. И.** Радиотехнические цепи и сигналы : учебник для вузов / С. И. Баскаков. – Москва : Высшая школа, 2016. – 462 с.

2 **Гаврилов, М. В.** Информатика и информационные технологии: учебник для вузов / М. В. Гаврилов. – Москва : Гардарики, 2017. – 655 с.

3 **Соловьев, И. В.** Проектирование информационных систем. Фундаментальный курс : учебное пособие для вузов / И. В. Соловьев, А. А. Майоров ; под ред. В. П. Савиных. – Москва : Академический Проект, 2016. – 398 с.

4 **Мельников, В. П.** Информационные технологии : учебник для вузов / В. П. Мельников. – 2-е изд., стер. – Москва : Академия, 2015. – 425 с.