

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное
учреждение высшего образования
«СЕВАСТОПОЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

ИНСТИТУТ ЯДЕРНОЙ ЭНЕРГИИ И ПРОМЫШЛЕННОСТИ

Кафедра систем контроля и управления АС

КОДИРОВАНИЕ ИНФОРМАЦИИ

Учебно-методическое пособие
для выполнения расчетно-графической работы
по дисциплине «Теория информации»

Севастополь
СевГУ
2019

УДК 004.9 (07)

К12

Рецензенты:

А.В. Углов – доцент кафедры ЭЭС, к.т.н., доцент;

К.Б. Матузаев – зав. кафедрой ЯЭУ, к.т.н., доцент

Качур С.А.

К12 Кодирование информации: учеб.-метод. пособие / С.А. Качур, А.А. Скидан, Н.В. Шахова. – Севастополь: СевГУ, 2019. – 16 с.; ил.

Рассмотрены различные методы кодирования информации.

Пособие предназначено для студентов направления подготовки 14.05.02 «Атомные станции: проектирование, эксплуатация и инжиниринг» очной и заочной форм обучения.

УДК 004.9 (07)

Рассмотрены и одобрены на заседании кафедры «Системы контроля и управления АС» ИЯЭиП СевГУ, протокол № 2 от 11 сентября 2019 г.

© Качур С.А., Скидан А.А., Шахова Н.В., 2019

© ФГАОУ ВО «Севастопольский
государственный университет», 2019

ОГЛАВЛЕНИЕ

1. Содержание и последовательность выполнения работы	4
2. Общие положения	5
3. Оптимальное кодирование	7
3.1. Понятия и принципы оптимального кодирования	7
3.2. Построение оптимального кода по методу Шеннона – Фано	9
3.3. Метод построения оптимальных кодов Хаффмена	9
4. Содержание пояснительной записки РГР	11
5. Контрольные вопросы	11
Литература	12
Приложение А. Международный код Бодо	13
Приложение Б. Пример выполнения работы	14

1. СОДЕРЖАНИЕ И ПОСЛЕДОВАТЕЛЬНОСТЬ ВЫПОЛНЕНИЯ РАБОТЫ

Тема: Оценка информационных характеристик текстового сообщения при кодировании равномерными и неравномерными кодами.

Цель:

1. Получить практические навыки в кодировании сообщений равномерными и неравномерными кодами.
2. Научиться кодировать текстовые сообщения по методам Шеннона – Фано и Хаффмена.
3. Научиться оценивать избыточность источника информации.

Задание: закодировать текстовое сообщение, состоящее из букв русского алфавита и включающее анкетные данные студента (фамилия, имя, отчество, название и адрес учебного заведения, группа) в коде Бодо, методами Шеннона – Фано и Хаффмена, оценить эффективность кодирования.

В результате выполнения расчетно-графической работы студент должен:

- **знать** методы кодирования информации;
- **уметь** получать равномерные и оптимальные неравномерные коды, проводить их оценку.

Расчетно-графическую работу рекомендуется выполнять **в следующей последовательности:**

1. Изучить общие теоретические сведения об оценке информационных характеристик текстовых сообщений и ознакомиться с алгоритмами кодирования Шеннона – Фано и Хаффмена.
2. Закодировать исходное сообщение равномерным кодом Бодо.
3. Закодировать это же сообщение по методу Шеннона – Фано.
4. Закодировать сообщение исходного текста по алгоритму Хаффмена.
5. Определить информационные характеристики текстового сообщения, оценить эффективность применения неравномерных кодов.
6. Оформить отчет.

2. ОБЩИЕ ПОЛОЖЕНИЯ

В общем случае **кодированием** называется отображение состояния одной физической системы «X» с помощью состояния некоторой другой «Y».

Наиболее простым случаем является ситуация, когда и отображающая и отображаемая системы имеют конечное число возможных состояний m и n соответственно.

Если число состояний системы «Y» меньше числа состояний «X», ($m < n$), то одно состояние системы «X» приходится отображать с помощью определенной последовательности (комбинации) состояний систему «Y».

Например, 33 буквы русского или 26 латинского алфавита отображаются четырьмя элементарными символами кода Морзе (точка, тире, короткая пауза, длинная пауза). Минимальное количество символов, состояний, которое может реализовать информационное устройство равно двум (1,0), такие коды называются **двоичными**.

Символы, при помощи которых записывается передаваемое (исходное) сообщение, составляют первичный алфавит, а символы последовых преобразований по определенным правилам – вторичный алфавит.

Коды, в которых сообщения представлены комбинациями с равным количеством символов, называются **равномерными** или **комплектными**.

Примером равномерного двоичного кода может служить широко применяемый в связи и вычислительной технике буквенно-цифровой код Бодо. Первые варианты кода Бодо были пятизначными, т.е. общее число комбинаций такого кода N не превышало 32:

$$N = 2^k = 32 \text{ (при } k = 5\text{)}. \quad (1)$$

По мере развития и совершенствования вычислительной техники усложнение равномерных буквенно-цифровых кодов шло в основном по пути добавления ряда служебных символов, используемых для расширения числа комбинаций и контроля правильности передаваемых сообщений.

Из теории информации известно, что максимальное количество информации на символ сообщения можно получить только в случае равновероятных и независимых символов:

$$H_{\max} = \log m. \quad (2)$$

Однако реальные коды с $m > 2$ редко удовлетворяют этому условию из-за наличия связей между символами, накладываемыми правилами составления (грамматикой) кодовых комбинаций. С этой точки

зрения энтропия может быть определена как информационная нагрузка на символ сообщения.

При неравном распределении вероятностей появления символов алфавита в тексте сообщения (например, в русском языке наиболее часто встречается буква «о», а в английском – «е») можно количественно оценить такие информационные характеристики как избыточность и коэффициент сжатия информации (или относительную энтропию).

Избыточность характеризует информационную недогруженность символов и оценивается выражением

$$D = 1 - \frac{H}{H_{\max}}, \quad (3)$$

где H – энтропия передаваемого сообщения с неравномерным распределением вероятностей появления символов в тексте; $H_{\max}$ – максимальная информационная нагрузка на символ при равномерном распределении.

$H_{\max}$ определяется достаточно легко при известном количестве символов в алфавите; так для русского языка с 32 буквами:

$$H_{\max} = \log_2 32 = 4,42$$

и для английского

$$H_{\max} = \log_2 26 = 4,7.$$

Подсчет H достаточно сложен и требует больших затрат. Поэтому можно использовать приближенное выражение

$$D = 1 - \frac{n_{\min}}{n}. \quad (4)$$

где $n_{\min}$ – теоретически возможное число элементарных символов; n – использованное число элементов кода для передачи сообщения.

Отношение информационной нагрузки на символ сообщения H к максимально возможной $H_{\max}$ в выражении (3) дает принципиальную возможность сжатия сообщения и называется **коэффициентом сжатия или относительной энтропией**.

$$\mu = \frac{H}{H_{\max}}. \quad (5)$$

Соответственно отношение $n_{\min}/n$ также характеризует коэффициент сжатия информации.

Избыточность D и коэффициент сжатия μ связаны между собой выражением:

$$D = 1 - \mu. \quad (6)$$

Устраняется статистическая избыточность путем построения эффективных (наиболее выгодных) кодированных сообщений. Общее правило неизвестно, но при решении этой задачи стараются составить сообщение таким образом, чтобы каждой букве соответствовала бы определенная комбинация символов «0» и «1» и чтобы среднее число этих символов на букву исходного текста было минимальным.

3. ОПТИМАЛЬНОЕ КОДИРОВАНИЕ

3.1. Понятия и принципы оптимального кодирования

Оптимальным кодированием называется процедура символов первичного алфавита m_1 в кодовые слова во вторичном алфавите m_2 , при котором средняя длина сообщения во вторичном алфавите имеет минимально возможную для данного m_2 длину.

Оптимальными называются коды, представляющие кодируемые понятия кодовыми словами минимальной средней длины.

Из основной теоремы кодирования вытекают **основные свойства** оптимальных кодов:

- минимальная средняя длина кодового слова оптимального кода обеспечивается в том случае, когда избыточность каждого кодового слова сведена к минимуму (в идеальном случае — к нулю);
- кодовые слова оптимального кода должны строиться из равновероятных и взаимонезависимых символов.

Из свойств оптимальных кодов вытекают принципы их построения.

Первый принцип оптимального кодирования: выбор каждого кодового слова необходимо производить так, чтобы содержащееся в нем количество информации было максимальным.

Второй принцип оптимального кодирования заключается в том, что буквам первичного алфавита, имеющим большую вероятность, присваиваются более короткие кодовые слова во вторичном алфавите.

Принципы оптимального кодирования определяют методику построения оптимальных кодов.

Коды, представляющие первичные алфавиты с неравномерным распределением символов, имеющие минимальную среднюю длину кодового слова во вторичном алфавите, называются **оптимальными неравномерными кодами (ОНК)**.

Максимально эффективными будут те ОНК, у которых

$$\log_2 m \sum_{i=1}^N l_i p_i = l_{cp} = H, \quad (7)$$

где m и N – символы соответственно вторичного и первичного алфавитов.

Длина двоичных кодов

$$l_{cp} = \sum_{i=1}^N l_i p_i = - \sum_{i=1}^N p_i \log_2 p_i, \quad (8)$$

$$l_i = -\log_2 p_i. \quad (9)$$

Эффективность ОНК оценивают при помощи:

1) **коэффициента статистического сжатия**, который характеризует уменьшение количества двоичных знаков на символ сообщения при применения ОНК по сравнению с применением методов нестатистического кодирования

$$K_{CC} = \frac{H_{\max}}{l_{cp}} = \frac{\log_2 N}{\log_2 m \sum_{i=1}^N p_i l_i}. \quad (10)$$

2) **коэффициента относительной эффективности**, который показывает, насколько используется статистическая избыточность передаваемого сообщения

$$K_{OЭ} = \frac{H}{l_{cp}}. \quad (11)$$

Для наиболее общего случая

$$K_{OЭ} = \frac{- \sum_{i=1}^N p_i \log_2 p_i}{\log_2 m \sum_{i=1}^N p_i l_i}. \quad (12)$$

3.2. Построение оптимального кода по методу Шеннона – Фано

Методика Шеннона – Фано для ансамбля из M сообщений сводится к следующему

1) множество из M сообщений располагают в порядке убывания вероятностей;

2) первоначальный ансамбль кодируемых сигналов разбивают на две группы таким образом, чтобы суммарные вероятности сообщений обеих групп были по возможности равны;

3) первой группе присваивают символ 0, второй группе – символ 1;

4) каждую из подгрупп делят на две группы так, чтобы их суммарные вероятности были по возможности равны;

5) первым подгруппам каждой из групп вновь присваивают 0, а вторым – 1, в результате чего получают вторые цифры кода. Затем каждую из четырех подгрупп вновь делят на равные (с точки зрения суммарной вероятности) части и т. д. до тех пор, пока в каждой из подгрупп останется по одной букве.

3.3. Метод построения оптимальных кодов Хаффмена

В методе Хаффмена используются для построения ОНК кодовые деревья.

Хаффмен показал, что для получения минимально возможной длины кода основанием m с число взаимонезависимых букв первичного алфавита N

$$l_{cp} = \sum_{i=1}^N l_i p_i, \quad (13)$$

необходимо и достаточно выполнение следующих условий:

1) если выписать символы в порядке убывания вероятностей

$$p_i \geq p_j, \text{ то при } i < j \quad l_i \leq l_j;$$

2) два последних, не больше чем m , кодовых слова равны по длине и отличаются лишь значениями последнего символа, при этом

$$2 \leq n_0 \leq m,$$

где m – число качественных признаков вторичного алфавита, а n_0 – число наименее вероятных сообщений, объединенных на первом этапе построения кодового дерева, кроме того

$$\frac{N - n_0}{m - 1} = a, \quad (14)$$

где a – целое положительное число;

3) любая возможная последовательность $l_N - 1$ кодовых слов должна быть кодовой комбинацией, либо иметь своим префиксом разрешенную комбинацию.

Исходя из данных условий, Хаффмен предложил следующий метод построения ОНК.

1) Символы первичного алфавита выписываются в порядке убывания вероятностей.

2) Последние n_0 символов, где $2 \leq n_0 \leq m$ и $(N - n_0)/(m - 1)$ – целое число, объединяют в некоторый новый символ с вероятностью, равной сумме вероятностей объединенных символов.

3) Последние символы с учетом образованного символа вновь объединяют и получают новый вспомогательный символ.

4) Опять выписывают символы в порядке убывания вероятностей с учетом вспомогательного символа т.д. до тех пор, пока вероятности m оставшихся после $(N - n_0)/(m - 1)$ -го выписывания не дадут в сумме 1.

На практике данную процедуру заменяют построением кодового дерева. Для кодов с числом качественных признаков $m = 2$ это сводится к тому, что символы кодируемого алфавита попарно объединяются в новые символы, начиная с символов, имеющих наименьшую вероятность, а затем, с учетом вероятностей вновь образованных символов опять производят попарное объединение символов с наименьшими вероятностями и таким образом строят двоичное кодовое дерево, в вершине которого стоит символ с вероятностью 1.

При построении ОНК для вторичных алфавитов с $m = 2$ методы Шеннона – Фано и Хаффмена дают в большинстве случаев одинаковые результаты.

Преимущество метода Хаффмена сказывается при построении ОНК для вторичных алфавитов с $m > 2$. Большая точность достигается за счет более строгого выбора числа наименее вероятных символов первичного алфавита, объединяемых на первом этапе построения кодового дерева.

Преимущество оптимальных кодов: при прочих равных условиях оптимальные коды позволяют вести передачу информации с максимальной скоростью.

Недостаток оптимальных кодов: при прочих равных условиях оптимальные коды наиболее беззащитны от влияния помех, т. е. обладают наименьшей помехоустойчивостью.

4. СОДЕРЖАНИЕ ПОЯСНИТЕЛЬНОЙ ЗАПИСКИ РГР

Пояснительная записка выполняется на листах формата А4 и включает следующее:

1. Титульный лист.
2. Задание на РГР в соответствии со своими анкетными данными.
3. Количественная оценка исходного текста.
4. Кодирование равномерным кодом Бодо.
5. Кодирование неравномерным кодом Шеннона – Фано.
6. Кодирование неравномерным кодом Хаффмена.
7. Оценка избыточности сообщения.
8. Перечень использованной литературы.

Примечание. Текстовое сообщение не должно содержать знаков препинания и цифр. Пробелы учитываются при рассмотрении. Пример выполнения РГР приведен в Приложении Б. Код Бодо представлен в Приложении А.

5. КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Определение кода.
2. Определение кодирования
3. Определение кодового слова
4. Определение декодирования
5. Понятие первичного алфавита
6. Понятие вторичного алфавита
7. Определение равномерного кода
8. Определение неравномерного кода
9. Избыточность сообщения
10. Коэффициент сжатия сообщения
11. Избыточность, вызванная статистической связью между символами сообщения
12. Избыточность, вызванная неравномерностью распределения символов в сообщениях
13. Помехоустойчивость системы передачи информации
14. Эффективность системы передачи информации
15. Надежность системы передачи информации

ЛИТЕРАТУРА

1. Вентцель Е.С. Теория вероятностей (Гл. 18. Теория информации) – Мю: Физматгиз, 1962.
2. Дмитриев В.Н. Прикладная теория информации. – Мю: Высшая школа, 1989.
3. Коган И.М. Прикладная теория информации. – Мю: Радио и связь, 1981.
4. Шеннон К. Работа по теории информации и кибернетике. Пер. с англ. – М.: Ил, 1964.
5. Орлов В.А., Филиппов Л.И. Теория информации в упражнениях и задачах. – Мю: ВШ, 1976.
6. Шишкевич Е. В. Методические рекомендации по дисциплине "Теория информации" для студентов заочной формы обучения. – Севастополь: СИЯЭиП, 2001.

Приложение А

Международный код Бодо

Кодовая комбинация	Значение кодовой комбинации			Кодовая комбинация	Значение кодовой комбинации		
	1-м	2-м	3-м		1-м	2-м	3-м
10 000	I	1	А	00 111	R	—	Р
00 110	B	8	Б	00 101	S	.	С
01 101	W	C	В	10 101	T	4	Т
01 010	G	7	Г	10 100	U	4	У
11 110	O	0	Д	01 100	F	9	Ф
01 000	E	2	Е	11 010	H	+	Х
00 010	V	Z	Ж	10 110	C	9	Ц
11 001	Z	:	З	10 111	Q	I	Ц
01 100	J	III	И	01 001	X	.	Ь
10 010	J	6	Й	00 100	Y	3	Ы
10 011	K	9	К	00 011	B	0	Я
11 011	L	II	Л	11 111			
01 011	M	S	М	00 010	Буквы русские		
01 111	N	Ю	Н	00 001	Цифры		
11 100	O	5	О	10 001	Буквы латинские		
11 000	P	T	П	00 000	Пробел		
					Звонок		

Приложение Б

Пример выполнения работы

Задание: выполнить кодирование кодом Бодо, методом Шеннона – Фано и Хаффмена следующий текст:

ИВАНОВ ИВАН ИВАНОВИЧ

Рассмотрим сообщение без учета пробелов.

Кодирование кодом Бодо

Буква	Частота появления	Вероятность p_i	Код
В	5	0,28	01 101
И	4	0,22	01 100
А	3	0,17	10 000
Н	3	0,17	01 111
О	2	0,11	11 100
Ч	1	0,05	10 110
Сумма	18	1	

Кодирование методом Шеннона – Фано

Буква	p_i	Код				l_i	$l_i p_i$
В	0,28	0	0	0	0	2	0,56
И	0,22	0	1			2	0,44
А	0,17	1	0			2	0,34
Н	0,17	1	1			3	0,34
О	0,11	1	1	1	0	4	0,44
Ч	0,05	1	1	1	1	4	0,2
Сумма	1						2,32

Кодирование методом Хаффмена

Построение кодового дерева

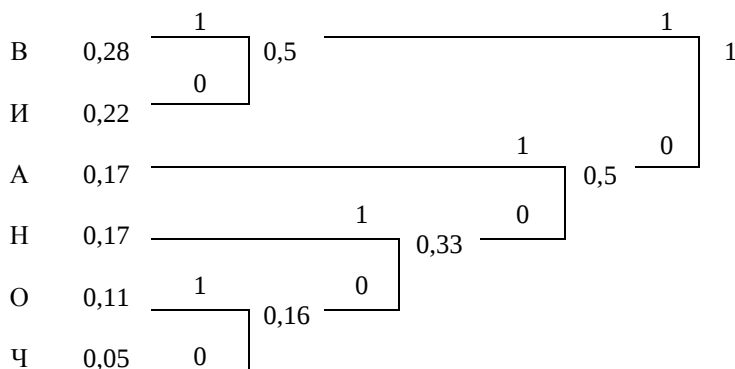


Таблица результатов кодирования

Буква	p_i	Код	l_i	$l_i p_i$
В	0,28	11	2	0,56
И	0,22	10	2	0,44
А	0,17	01	2	0,34
Н	0,17	001	3	0,34
О	0,11	0001	4	0,44
Ч	0,05	0000	4	0,2
Сумма	1			2,32

Сравнение результатов кодирования различными методами

По результатам количественной оценки сообщения имеем сообщение из 18 символов первичного алфавита без учета пробела, полученные результаты сводим в таблицу:

Код	Первич- ный ал- фавит	Вторич- ный ал- фавит	Средняя длина символа в сообще- нии, l_{cp}	Избыточность, D
Бодо	18	90	$\frac{90}{18} = 5$	$D = 1 - \frac{5}{5} = 0$ нет сжатия
Шен- нона - Фэно	18	55	$\frac{55}{90} = 3,06$	$D = 1 - \frac{3,06}{5} = 0,39$ сжатие 0,61
Хаф- фмена	18	55	$\frac{55}{90} = 3,06$	$D = 1 - \frac{3,06}{5} = 0,39$ сжатие 0,61

Таким образом, в сообщении длиной 18 символов избыточность при кодировании кодом Бодо отсутствует, сжатия сообщения нет. Кодирование как неравномерным кодом Шеннона – Фэно, так и неравномерным кодом Хаффмена дает избыточность равную 0,39, сжатие 0,61.

Издается в авторской редакции