DOI: 10.14451/1.178.107

ВОЗМОЖНЫЕ ПРЕДСТАВЛЕНИЯ О КЛАССАХ КОДОВ И АЛГОРИТМАХ КОДИРОВАНИЯ

© 2019 **Данелян Тэя Яновна**
кандидат экономических наук,
доцент кафедры Прикладной информатики и информационной безопасности
РЭУ им. Г.В. Плеханова, Россия, Москва
E-mail: tdanelan@yandex.ru

© 2019 **Епихин Максим Николаевич**
научный сотрудник факультета Математической экономики, статистики и информатики, направления «ИиВТ»
РЭУ им. Г.В. Плеханова, Россия, Москва
E-mail: mepihin@yandex.ru.

© 2019 **Щукин Андрей Дмитриевич**
научный сотрудник факультета Радио и телевидения
Московский технический университет связи и информатики (МТУСиИ), Россия, Москва
E-mail: A.pike2110@yandex.ru.

В статье рассмотрены основные положения, связанные с классами кодов, кодированием, теорией криптографии, практическими примерами реализации методов кодирования (криптографии).

Ключевые слова: классы кодов, методы (алгоритмы) кодирования, информация, информационные системы, коды, кодирование

Введение

Разработчики информационных систем, которые используются в различных направлениях человеческой деятельности, должны знать системы кодирования, классы кодов, алгоритмы и методы кодирования с тем, чтобы уметь правильно организовывать как защиту информации и безопасность информации, так и хранение информации на любых носителях и в любой цифровой среде. Именно поэтому все направления, связанные с кодированием информации являются интересными и актуальными, а любые предложения по способам кодирования, способам хранения информации являются первостепенно важными при создании информационных систем.

В статье предлагается рассмотреть некоторые способы кодирования информации, которые минимизируют память хранения информации и время передачи информации. В статье также приводятся основные положения кодирования, классификация кодов и сами алгоритмы оптимального кодирования информации.

Сущность криптографии

Теория криптографии связана с теорией информации, кодирования и кодами.

Теория кодирования (криптография) — раздел теории информации, изучающий коды, отображающие сообщения (информационные объекты) заданного вида в слова из символов некоторого заданного алфавита.

Таким образом код — универсальный способ отображения информации при ее хранении, передачи и обработке в виде системы соответствий между элементами сообщений (информационные объекты) и сигналами, при помощи которых эти элементы (информационные объекты) можно зафиксировать.

Если код — это слово информационного языка, представляющее собой производную единицу, которая состоит из более простых единиц, семантические множителей, и которые обеспечивают аналитическое задание парадигматических отношений, функциональную полноту СС [6; 7; 8], то кодирование — процесс структурного синтеза информационных объектов, в ре-

зультате которого осуществляется кодирование состояний информационного объекта набора состояний его элементов [6; 7; 8]. Понятие кода, процесса и инструментария кодирования составляет сущность криптографии, которая в свою очередь является частью теории информации (ТИ).

Различные виды кодов применяются для представления дискретной информации в информационных системах (в линиях, каналах связи, системах автоматики, вычислительных устройствах). То есть информация отображается в информационных системах в виде кодов.

Пусть дано: $B = \{b_i\}$ множество сообщений и некоторый алфавит $A = \{a_i\}^m$, ($a_i \in A$,символ) ($b_i \in B$-множеству сообщений) Конечная {совокупность$\langle a_i \rangle$} это слово в алфавите.

Следовательно, множество слов в А будет называется кодом, если код поставлен во взаимно однозначное соответствие с множеством В. Каждое слово, входящее в код, называется кодовым словом (кодовой комбинацией) $b_i \leftrightarrow \langle a_{n-1} .. a_1, a_0 \rangle$.

Число символов в кодовом слове $\langle a_{n-1} .. a_0 \rangle$ называется длиной кода (кодового слова). Для записи кодового символа a_i используются различные обозначения — цифры, буквы, специальные знаки. Число различных значений $\langle m \rangle$, которые может принимать каждый кодовый символ a_j, называется основанием кода. Тогда кодовое слово $k = \langle a_{n-1}, a_{n-2} .. a_i, .., a_0 \rangle$ имеет длину n и называется n-разрядным числом в системе счисления с основанием m:

$$k = \sum_{i=n-1}^{0} a_i m^i = a_{n-1} m^{n-1} + .. + a_1 m^0 1 + a_n ,$$

где $m^0 = 1$

Существуют общепринятые способы кодирования, а именно коды Шенона, статистические коды, составление которых реализуется после определения длины кода.

Помехоустойчивые коды — разделимые и неразделимые, групповые и некритические. Если код построен так, что имеет место взаимное соответствие с множеством сообщений, то для каждого сообщения будет иметь место своя кодовая комбинация.

Выбор кодового слова для каждого слова зависит от системы кодирования и выражает парадигму сообщений. Для записи кодовой комбинации могут использоваться разные символы. Число различных значений М, которые может принимать кодовый символ называется основанием данного кода. Если длина кодового слова М, то имеет место М разрядное кодовое слово. Кодовые слова могут иметь одинаковую или различную длину.

В целях минимизации длины кодового слова применяются статистические коды Шенона и Хаффмена, а также азбука Морзе, префиксные коды. Основанием кодового слова в равномерном коде чаще всего является основание 2. Что предопределяется видом информационной системы обработки информации, которая базируется на двоичной системе кодирования.

Классы кодов бывают:

1. Неравномерные
2. Равномерные
3. Коды Шенона
4. Помехоустойчивые коды
5. Симметричные и асимметричные

На рисунке 1 приведено дерево классификации кодов.

Неравномерные коды применяются в системах кодирования, где учитываются статистические свойства сообщений для минимизации средней длины кодового слова на элемент сообщения. Имеют место эффективные методы кодирования информации, которые учитывают статистическую структуру сообщений (код Шеннона, код Харфмена). Известные коды — код Морзе для кодирования алфавитно-цифровой информации, префиксные коды без разделительных знаков между кодовыми словами. Префиксные коды имеют свойство самосинхронизации. С помощью этого свойства однозначно разделяются кодовые слова в последовательности сообщений.

Равномерные коды. Основанием кодового слова в равномерном коде чаще всего является число два (2), то есть имеет место двоичная система кодирования (двоичные коды). Двоичная система кодирования предопределяется системой обработки информации, которая основана на двух устойчивых состояниях (есть — нет сигнала или 1–0).

Двоичные позиционные коды:

$2^{n-1}, 2^{n-2}, .., 2^i, 2^0$, где n — число разрядов

В равномерном кодировании определяются два подкласса: прямой и обратный.

Непозиционные коды (символические), реф-

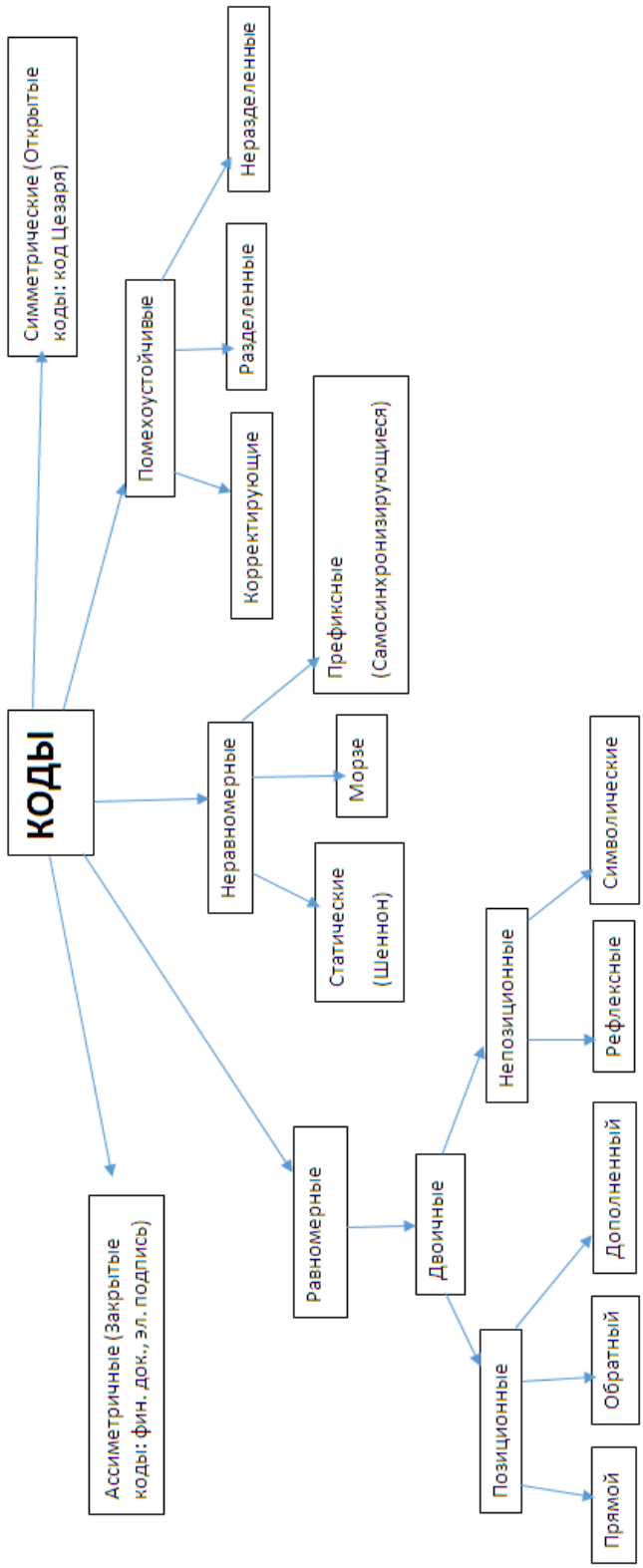


Рис. 1. **Топологическая схема дерева классификации кодов**

лексные.

Коды К. Шеннона — статистические коды математическое ожидание длины кода, которых вычисляется как математическое ожидание от достоверности кода длиной n: E(n(K)) = nH, где H — энтропия, то есть степень потери точности кодового слова K, и n — длина кодового слова K.

Помехоустойчивые коды:

1. Корректирующие коды
2. Разделимые/Неразделимые коды
3. Групповые
4. Арифметические

Симметричные коды, или открытые коды (ОК) — их называют кодом Цезаря.

Асимметричные коды, или закрытые коды (ЗК) коды типа (5) и (6) используются при кодировании финансовых документов с электронной подписью.

Алгоритмы кодирования

Ниже представлены алгоритмы кодирования и декодирования в среде алгоритмического языка «упрощенный Алгол» на рис. 2–5.

Input: Объем алфавита M, вероятности букв
Output: Двоичное дерево кода Хаффмена

Инициализация:
количество необработанных узлов $M_0 = M$
while $M_0 > 1$ **do**
В списке необработанных узлов найти два узла с наименьшими вероятностями.
Исключить эти узлы из списка необработанных.
Ввести новый узел, приписать ему суммарную вероятность двух исключенных узлов.
Новый узел связать ребрами с исключенными узлами.
$M_0 \leftarrow M_0 - 1$.
end

Рис. 2. **Алгоритм 1**

Input: Объем алфавита M
вероятности букв $p_i, i = 1, ..., M$
длина последовательности n
последовательность на выходе источника $(x_1, ..., x_n)$,
Output: Кодовое слово арифметического кода

Кумулятивные вероятности:
$q_1 = 0$;
for $i = 2$ **to** M **do**
$q_i = q_{i-1} + p_{i-1}$;
end

Кодирование:
for $i = 1$ **to** n **do**
$F \leftarrow F + q(x_i)G$;
$G \leftarrow p(x_i)G$;.
end

Формирование кодового слова:
$\boldsymbol{c} \leftarrow$ первые $\lceil -\log G \rceil + 1$ разрядов после запятой в двоичной записи числа $F + G/2$.

Рис. 3. **Алгоритм 2**

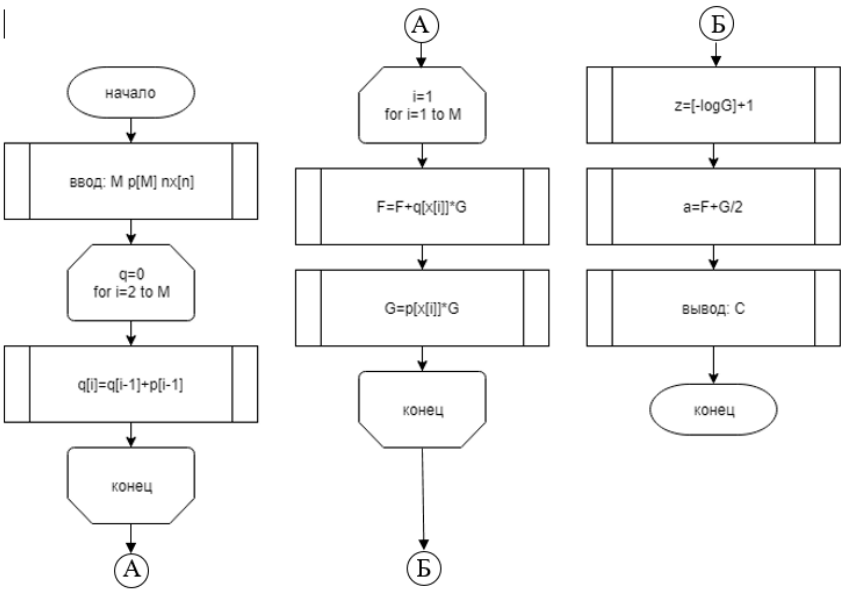


Рис. 4. **Блок схема к алгоритму 2**

Input: Объем алфавита M
вероятности букв $\{p_1, ..., p_M\}$
кумулятивные вероятности букв $q_i, i = 1, ..., M$
длина декодируемой последовательности n
кодовое слово в виде числа $\hat{F}$.
Output: Декодированная последовательность букв $(x_1, ..., x_n)$

Инициализация: $q_{M+1} = 1$; $S = 0$; $G = 1$.

Декодирование:
for $i = 1$ **to** n **do**
  $j = 1$;
  while $S + q_{j+1}G < \hat{F}$ **do**
    $j \leftarrow j + 1$.
  end
  $S \leftarrow S + q_jG$;
  $G \leftarrow p_jG$;
  $x_i = j$.
end

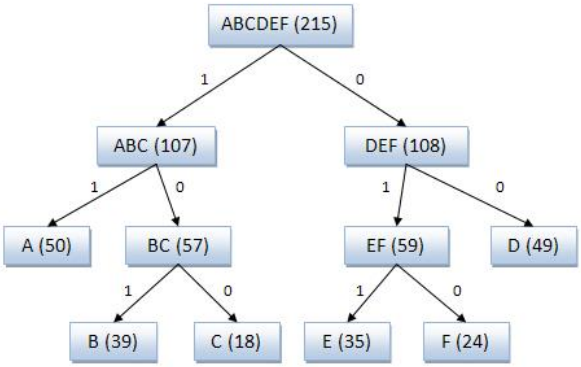
Результат:последовательность $(x_1, ..., x_n)$;

Рис. 5. **Алгоритм 3**

Алгоритм кодирования по Шеннону

Рассмотрим пример. Исходные символы:

1. A (частота встречаемости 50)
2. B (частота встречаемости 39)
3. C (частота встречаемости 18)
4. D (частота встречаемости 49)
5. E (частота встречаемости 35)
6. F (частота встречаемости 24)



Полученный код: A — 11, B — 101, C — 100, D — 00, E — 011, F — 010.

Алгоритм кодирования по Хаффмену

Сжатием информации в памяти компьютера называют такое её преобразование, которое ведёт к сокращению объёма хранимой памяти при сохранении закодированного содержания. Рассмотрим один из способов сжатия текстовой информации — алгоритм Хаффмена. С помощью этого алгоритма строится двоичное дерево, которое позволяет однозначно декодировать двоичный код, состоящий из символьный кодов различной длины. На рисунке 6 приведён пример такого дерева, построенный для алфавита английского языка с учётом частоты встречаемости его букв.

Закодируем с помощью данного дерева слово «hello»:

0101 100 01111 01111 1110

При размещении этого кода в памяти побитово он примет вид: 01011000111101111111110

Таким образом, текст, занимающий в кодировки ASCII 5 байтов, в кодировке Хаффмена займет 3 байта.

Алгоритм построения кодового дерева кода Хаффмена

Input: Объем алфавита M, вероятности букв

Output: Двоичное дерево кода Хаффмена

Инициализация:

Количество необработанных узлов $M0 = M$

While $M0 > 1$ do

В списке необработанных узлов найти два узла с наименьшими вероятностями.

Исключить эти узлы из списка необработанных.

Ввести новый узел, приписать ему суммарную вероятность двух исключенных узлов.

Новый узел связать рёбрами с исключенными узлами. $M0 \leftarrow M0–1$. End

Методика кодирования Щукина и Епихина.

В основе метода кодирования стоит переход между строчками, т.к. изначально рассчитывалось кодировать слова меньшим количеством цифр. Для того чтобы уменьшить количество мы решили использовать по 5 символов на букву. Строчки будут иметь 28 символов алфавита и 2 символа будут служить переходными вверх и вниз по строкам. Набор «0000» и «1111» переносят нас на другую строчку алфавита. Алфавит

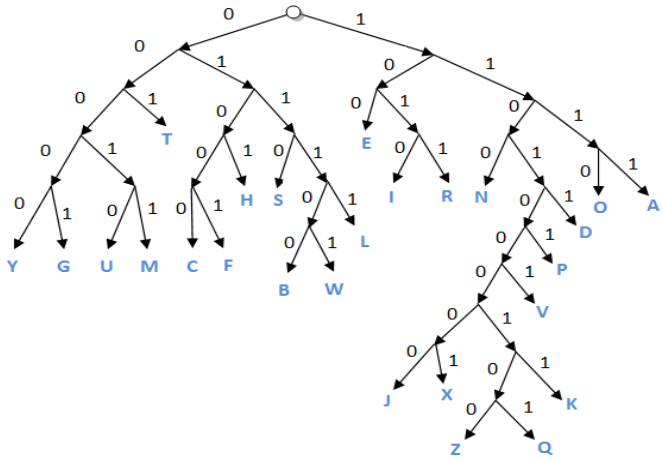


Рис. 6. **Пример дерева алфавита**

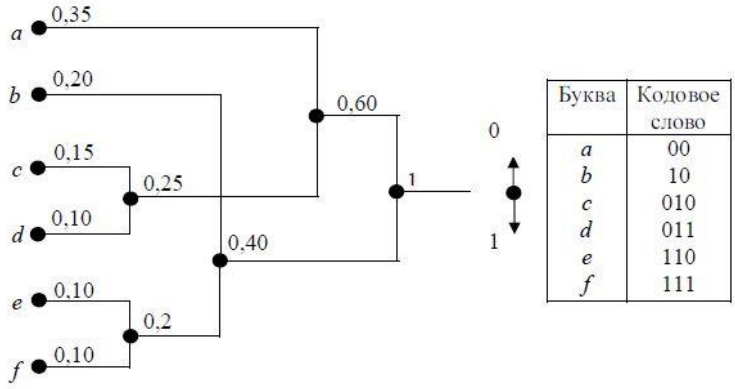


***Рис. 7*. Пример построения кода Хаффмена**

был составлен на основе открытых данных по частоте использования букв и символов в тексте. Алфавит будет выглядеть образом, представленным в таблице 1. Бинарные коды для символа имеют вид, представленным в таблице 2.

Расположение данного алфавита было сделано с использованием статистики частоты использования букв, для уменьшения количества переходов.

Тестирование показало, что данный алгоритм имеет преимущество на 32% по параметру «Отношение объема символов к методу кодирования» (таблица 3).

Алгоритм заключается в следующем:

1. Определяем наличие буквы и ее позицию.

2. При необходимости делаем переход вверх или вниз от текущей строки (первая строка по умолчанию)

3. Кодируем букву соответствующим бинарным кодом.

Для примера кодирования возьмем слово «Мозг»: в начале мы находимся на первой строке. На этой строке у нас находятся буквы «М» и «О». Их порядковые номера 0 и 11 соответственно. Значит кодируем их бинарными символами: 00001 и 01100 соответственно. Так как все буквы алфавита находятся на одной строке, то просто подставляем значения.

Слово «Щи» имеет уже переход на строку. Сначала у нас первая строка. Буква «Щ» находится на второй. Следовательно, надо сделать переход вниз «11111» затем взять букву по индексу «10001». Далее перейти на строку вверх «00000» и взять букву по индексу «00100». В итоге имеем слово: «11111100010000000100» [9].

Таблица 1. **Строки алфавита**

Строка 1	ОЕАИТНСРВЛКМДПУЯЫГЗБЧЙХЪЖЬЮ
Строка 2	0123456789.,:/ШЦЩЭФЁ+-?!()%=

Таблица 2. **Бинарные символы**

00001	00010	00011	00100	00101	00110	00111
01000	01001	01010	01011	01100	01101	01110
01111	10000	10001	10010	10011	10100	10101
10110	10111	11000	11001	11010	11011	11100

Таблица 3. **Отношение объема символов к методу кодирования**

Объем					
	5bit (КБ)	4bit (КБ)	8 bit (КБ)	% 4bit -> 8 bit	% 5bit -> 8 bit
2525	13,16	14,32	19,56	73%	67%
88673	249,59	271,81	371,12	73%	67%
354692	998,27	1087,17	1484,40	73%	67%

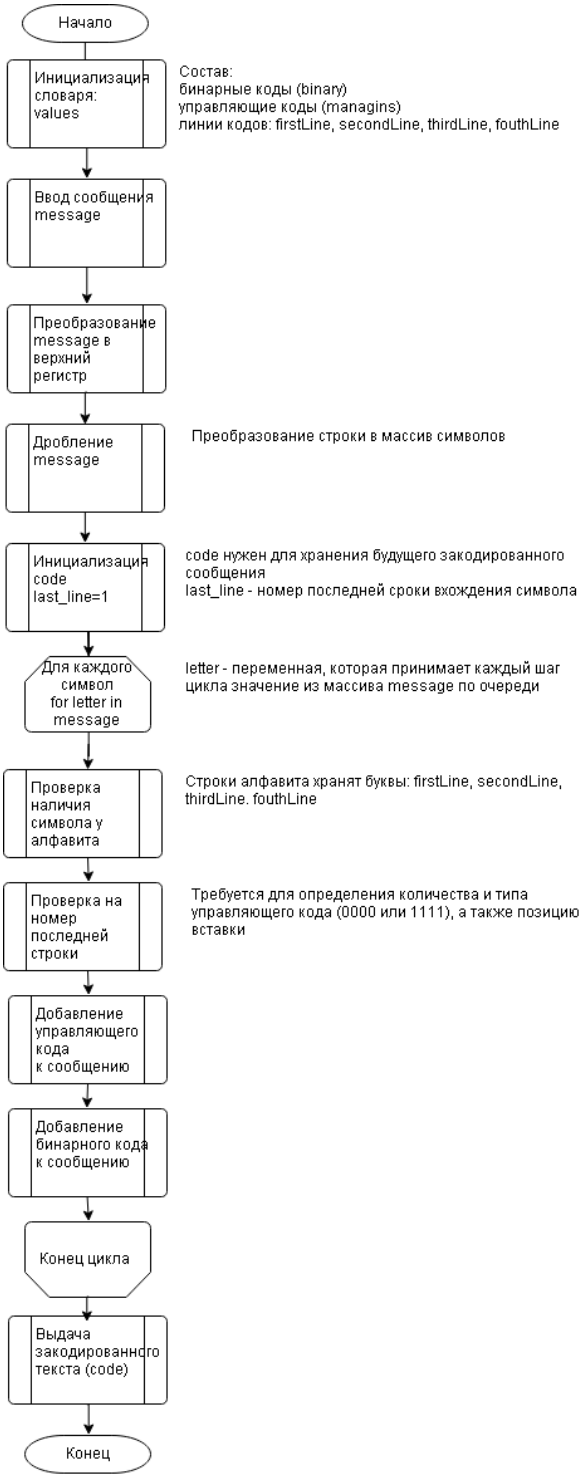


***Рис. 8*. Схема алгоритма функционирования**

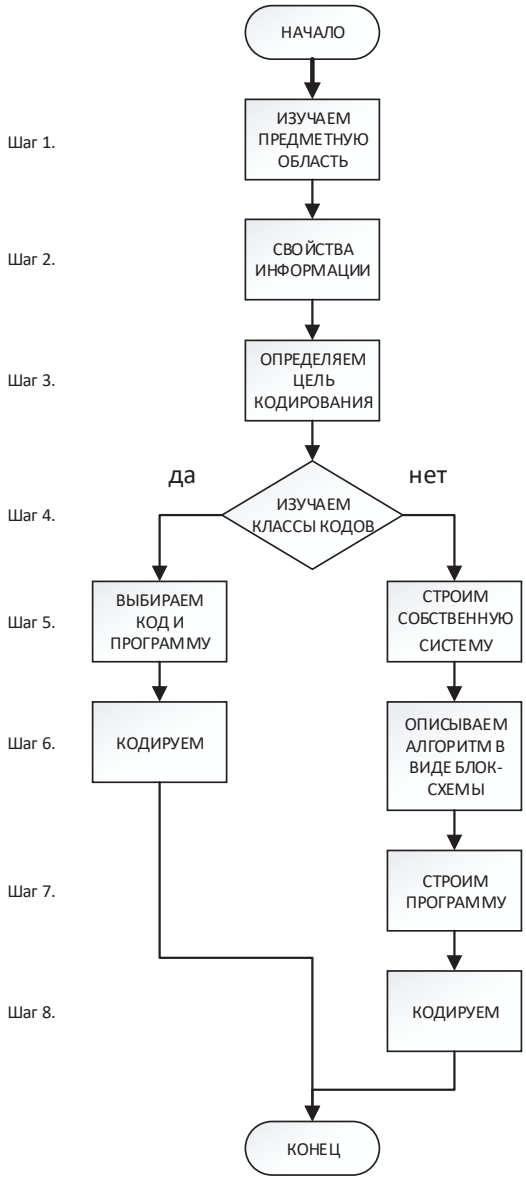


***Рис. 9*. Схема технологического процесса создания информационной системы кодирования**

Библиографический список

1. *Данелян Т.Я.* ТЕОРИЯ СИСТЕМ И СИСТЕМНЫЙ АНАЛИЗ (ТСиСА): учебно-методический комплекс / Т. Я. Данелян.— М.: Изд. центр ЕАОИ, 2010.— 303 с.
2. *Данелян Т.Я.* ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ В ПСИХОЛОГИИ: учебно-методический комплекс / Т. Я. Данелян.— М.: Изд. центр ЕАОИ, 2014.— 227 с.
3. *Данелян Т.Я.* Проектно-ориентированные ЭИС: Учебное пособие / Московский государственный университет экономики, статистики и информатики. М., 2013 — с.
4. *Данелян Т.Я*, *Квятковский А.В.* Информационные технологии в сфере юриспруденции / Российский экономический университет имени Г. В. Плеханова (РЭУ им. Г. В. Плеханова) — М., 2016 г.— 105с.
5. *Данелян Т.Я.* «Экономические информационные системы предприятий и организаций» (ч-1), Москва, 2005 г.
6. Общая теория информации (ОТИ). Учебно-методический комплекс / Т. Я. Данелян, М. Н. Епихин.— Москва: Русайнс, 2018.— 116 с.— ISBN978–5–4365–2869–4.
7. Энциклопедия кибернетики в двух томах (отв. ред. В. М. Глушков). Киев: Украинская советская энциклопедия, 1974.
8. Словарь по кибернетике (под ред. В. М. Глушкова). Киев: Главная редакция Украинской советской энциклопедии, 1979
9. 5 bit encoder URL: https://github.com/mepihindeveloper/encoder (дата обращения: 03.03.2019).