Кодирование информации

Оглавление

Краткие теоретические сведения

Для кодирования информации используются знаковые системы. Знаковая система состоит из набора знаков (символов), который называется **алфавитом**. Полное количество символов алфавита называется **мощностью алфавита**. Например, алфавит русского языка состоит из 33 букв, латинского – из 26 букв.

Минимально возможное количество символов в алфавите равно двум. Существующие технические электронные устройства надежно сохраняют и распознают только два различных состояния, поэтому именно такой алфавит используется в компьютере. Он называется двоичным алфавитом, его символы – цифры «0» и «1». С помощью этих двух символов можно представить любую информацию в компьютере.

Если для сообщения используется двоичный алфавит, и длина сообщения – один знак, можно составить два различных сообщения (0 и 1). Если длина сообщения – два знака, можно сформировать $2 \times 2 = 2^2 = 4$ разных комбинации (00, 01, 10, 11). При длине сообщения три знака получим $2 \times 2 \times 2 = 2^3 = 8$ различных комбинаций (000, 001, 010, 011, 100, 101, 110, 111) и т.д.

В информатике один знак двоичного алфавита называют битом (от **bi**nary digi**t** – двоичная цифра).

Кодовый алфавит и кодовое слово

Решать задачу кодирования информации человечество начало задолго до появления компьютеров: великие достижения человечества — письменность и арифметика — не что иное, как системы кодирования речи и числовой информации. Задачи кодирования решались и для передачи информации с помощью технических устройств.

Для хранения в компьютере и передачи информации по каналам связи символы должны быть закодированы при помощи некоторого кодового алфавита – набора знаков, при помощи которых можно составлять слова.

Каждый символ исходного алфавита (мощности N) при кодировании представляется последовательностью символов кодового алфавита (мощности M), которая называется **кодовым словом**[1]**. Код** состоит из кодовых слов.

Рассмотрим в качестве примера знакового кодирования азбуку Морзе.

1) В исходный алфавит входили буквы латинского алфавита, цифры, знаки препинания.

[1] Иногда кодовое слово называют кратко кодом.

2) Кодовый алфавит Морзе состоит из трех символов (*M*=3):

- тире – длинный сигнал,
- точка – короткий сигнал,
- пауза – отсутствие сигнала.

3) Каждый символ (знак) исходного алфавита Самуэль Морзе обозначил уникальной комбинацией из длинных и коротких сигналов – кодовым словом.

Кодовые слова однозначно определяли каждый символ исходного алфавита. Впоследствии к латинскому алфавиту добавились шифры для знаков национальных алфавитов, например, русского.

Принцип кодирования азбуки Морзе исходит из того, что буквы, которые чаще употребляются в английском языке, кодируются более короткими сочетаниями точек и тире. Это делает передачи компактнее, а такие коды называются **неравномерными**. Примеры кодов Морзе некоторых символов:

Символ исходного алфавита	Кодовое слово	Символ исходного алфавита	Кодовое слово	Символ исходного алфавита	Кодовое слово
A	• –	I	• •	J	• – – –
M	– –	L	• – • •	W	• – –

Заметим, что началом кодовых слов символов J, L, W является кодовое слово символа A. Поэтому невозможно однозначно декодировать полученное сообщение, если не использовать паузы между кодовыми словами. Например, требуется расшифровать сообщение, закодированное азбукой Морзе и переданное без пауз между кодами символов (используются только приведенные в таблице кодовые слова):

• – – – • – • •

Для декодирования сообщения будем последовательно слева направо выделять коды символов. Получим варианты декодирования: AMAI **(• – – – • – • •),** AML **(• – – – • – • •),** JAI **(• – – – • – ••),** JL **(• – – – • – • •).**

Для однозначного декодирования сообщения, закодированного азбукой Морзе, используют паузы, разделяющие кодовые слова.

Префиксные коды

Для того чтобы можно было однозначно декодировать сообщение, закодированное неравномерным кодом, без специального разделения кодов символов, используют так называемые префиксные коды.

Префиксный код – это код со словами переменной длины, в котором ни одно кодовое слово не является началом другого кодового слова.

Азбука Морзе – пример непрефиксного кода.

Пример префиксного двоичного кода для исходного алфавита из трех символов: 0, 10, 11. Сообщение 100111011010 однозначно декодируется: 10 0 11 10 11 0 10.

Набор 0, 10, 100, 11 образует непрефиксный код. Приведенное выше сообщения можно декодировать несколькими способами

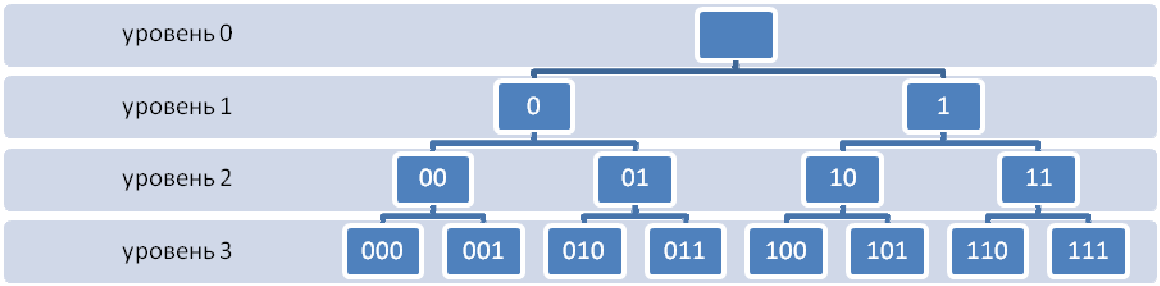
10 0 11 10 11 0 10

100 11 10 11 0 10

Сообщения, закодированные префиксными кодами, можно декодировать «на лету», не дожидаясь получения всего сообщения целиком. Префиксные коды используются для кодирования аудио и видео файлов, поэтому можно слушать музыку или смотреть видео до того, как файл загрузится целиком.

Для построения префиксного кода удобно использовать двоичное дерево. В двоичном дереве у каждого узла (кроме листьев) может быть один или два потомка, у каждого узла (кроме корня дерева) – ровно один родитель. Узлы, не имеющие потомков, называют листьями.

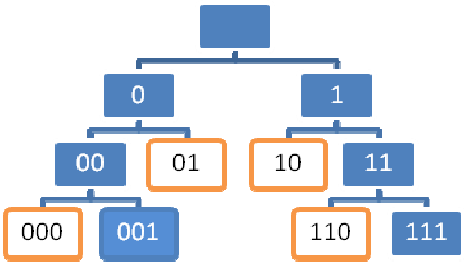
Корень дерева – пустая строка. В узлах дерева двоичные коды. Левые ветви дерева соответствует команде «приписать 0 справа», правые ветви – команде «приписать 1 справа».



Обратите внимание: в полученном дереве коды, расположенные на одном уровне дерева, образуют последовательность целых чисел, записанных в двоичной системе счисления.

Как построить префиксный код по кодовому дереву? Будем считать некоторые узлы кодовыми словами. Если узел выбран в качестве кодового слова, дальнейшие построения от этого узла не проводятся, узел не должен иметь потомков. Это обеспечивает соблюдение свойства префиксности кода – ни одно кодовое слово не является началом другого кодового слова. Таким образом, узлы, имеющие потомков, не могут быть кодовыми словами. **Кодовые слова образуют листья кодового дерева.**

На рисунке ниже листья выделены цветом. От узлов 001 и 111 можно продолжать построение дерева для получения новых кодовых слов.



Равномерные коды

Для того чтобы можно было однозначно прочитать сообщение, не разделяя коды символов специальными знаками, можно использовать равномерные коды. В этом случае кодовые слова всех символов исходного алфавита имеют одинаковую длину. Расшифровать такое сообщение не составит труда, но сообщение становится длиннее, чем при использовании неравномерных кодов.

Если мощность кодового алфавита равна M, длина кода – I, можно составить $N = M^I$ различных кодовых слов.

При использовании двоичного алфавита (M=2) для построения равномерного кода при длине кодового слова I количество различных кодовых слов N:

$$N = 2^I,$$

Таким образом, можно закодировать N символов исходного алфавита.

Текстовая информация состоит из букв, цифр, знаков препинания, специальных символов, таких как пробел, символ перевода строки и др. В случае, когда код каждого символа занимает в памяти компьютера 8 бит, общее количество символов, которые можно закодировать, равно $2^8 = 256$. Если кодовое слово состоит из 16 бит, можно закодировать $2^{16} = 65536$ символов.

Примеры решения заданий

Пример 1 задания с кратким ответом

Какой должна быть минимальная длина равномерного двоичного кода, если требуется составить 18 различных кодовых комбинаций?

Решение. Количество комбинаций есть символы исходного алфавита, которые кодируются двоичным кодом. Мощность двоичного кодового алфавита M=2, мощность исходного алфавита (количество различных комбинаций) N=18. Известно, что $N = 2^I$.

Определим длину двоичного кода $I = \log_2 N = \log_2 18$. Округлим полученный результат до ближайшего большего целого, получим I=5.

Ответ: 5

Пример 2 задания с кратким ответом

Световое табло состоит из лампочек. Каждая лампочка может находиться в одном из трех состояний («включено», «выключено» или «мигает»). Какое наименьшее количество лампочек должно находиться на табло, чтобы с его помощью можно было передать 18 различных сигналов?

Решение. Мощность кодового алфавита $M = 3$ (три состояния лампочек), мощность исходного алфавита $N = 18$ (количество различных сигналов). Известно, что $N = M^I = 3^I$. Определим длину кода по формуле $log_3 N = log_3 18$. Округлим полученный результат до ближайшего большего целого $2 < log_3 18 < 3$.

Ответ: 3

Пример 3 задания с кратким ответом

Для передачи сигналов на флоте используются специальные сигнальные флаги, вывешиваемые в одну линию (последовательность важна). Какое количество различных сигналов может передать корабль при помощи пяти сигнальных флагов, если на корабле имеются флаги трех различных видов (флагов каждого вида неограниченное количество)?

Решение. Мощность кодового алфавита (количество различных видов флагов) $M = 3$, длина кодового слова равна 5 (количество сигнальных флагов). Количество различных сигналов определим по формуле $N = M^I = 3^5 = 243$.

Эту задачу можно решить простыми рассуждениями. Так как имеем неограниченное количество флагов трех видов, то каждый флаг в последовательности из пяти сигнальных флагов можно выбрать тремя способами. Получаем 3 • 3 • 3 • 3 • 3 = 243.

Ответ: 243

Пример 4 задания с выбором одного ответа

Для 5 букв латинского алфавита заданы их двоичные коды (для некоторых букв – из двух бит, для некоторых – из трех). Эти коды представлены в таблице:

A	B	C	D	E
000	11	01	001	10

Определите сообщение в этой кодировке, которое может быть корректно декодировано (не содержит ошибки).

1) 11010001001001110

2) 110000000011011110

3) 11000001001111010

4) 11000000101111010

Решение:

Представленные коды являются префиксными, поэтому сообщение должно однозначно декодироваться. Попытаемся декодировать ответы, выделяя коды символов с начала строки:

1) 11 01 000 10 01 001 11 0

2) 11 000 000 001 10 11 11 0

3) 11 000 001 001 11 10 10

4) 11 000 000 10 11 11 01 0

В вариантах ответов 1, 2, 4 строка заканчивается кодовым словом 0, которого нет в заданном коде.

Ответ: № 3

Пример 5 задания с выбором одного ответа

Получено сообщение, переданное с использованием двоичных кодов:

1101110100011010000 10

В таблице представлены символы исходного алфавита и их коды

Л	И	Т	К	О
10	**111**	**110**	**010**	**00**

Сколько символов исходного алфавита содержит сообщение

1) 8 2)9 3) 10 4) 11

Решение:

Представленные коды являются префиксными, поэтому сообщение должно однозначно декодироваться. Будем выделять коды символов в сообщении с начала строки

110 111 010 00 110 10 00 010

Т И К О Т Л О К

В сообщении 8 символов.

Ответ: № 1

Решения заданий демоварианта 2012

Задание А9

Характеристики задания

Проверяемые элементы содержания	Умение кодировать и декодировать информацию
Контролируемый элемент содержания (по кодификатору)	1.1.2. Процесс передачи информации, источник и приемник информации. Сигнал, кодирование и декодирование. Искажение информации
Требования к уровню подготовки (по кодификатору)	1.1.3. Строить модели объектов, систем и процессов . Записывать алгоритмы на естественном языке и в виде блок - схем
Вид деятельности	Применение знаний и умений в новой ситуации
Уровень	базовый
Максимальный первичный балл	1
Время выполнения	2 мин.

Задание

А9 Для кодирования некоторой последовательности, состоящей из букв А, Б, В, Г и Д, решили использовать неравномерный двоичный код, позволяющий однозначно декодировать двоичную последовательность, появляющуюся на приёмной стороне канала связи. Использовали код: А–1, Б–000, В–001, Г–011. Укажите, каким кодовым словом должна быть закодирована буква Д.

Длина этого кодового слова должна быть наименьшей из всех возможных. Код должен удовлетворять свойству однозначного декодирования.

1) 00

2) 01

3) 11

4) 010

Решение

В задании используется префиксный код – код со словами переменной длины, в котором ни одно кодовое слово не является началом другого кодового слова. Он позволяет однозначно декодировать сообщение без специального разделения кодов символов.

Способ 1. Проверим, являются ли предложенные ответы началом заданных кодовых слов или наоборот, являются ли заданные коды началом кодов, предложенных в ответах.

1) 00 – является началом кода символа Б (000)

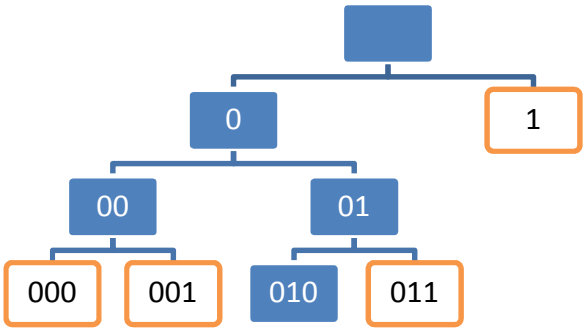
2) 01 – является началом кода символа Г (011)

3) 11 – код символа А (1) является началом этого кода

4) 010 – не является началом ни одного из заданных кодов символов, ни один из заданных кодов не является началом этого кода.

Таким образом, кодом буквы Д может быть только 010.

Способ 2. Пригодится, если задание перейдет в часть В. Построим кодовое дерево, в котором заданные коды должны быть листьями. Так как максимальная длина заданных в условии кодовых слов равна трем, потребуется не более четырех уровней двоичного дерева.



Из построенного дерева очевидно, что продолжать построение кодов можно лишь из узла 010. По условию задания длина кодового слова должна быть наименьшей из всех возможных, следовательно, это должен быть код 010.

Ответ: № 4