5 Тема: Кодирование и декодирование информации.

Что нужно знать:

- кодирование – это перевод информации с одного языка на другой (запись в другой системе символов, в другом алфавите)
- обычно кодированием называют перевод информации с «человеческого» языка на формальный, например, в двоичный код, а декодированием – обратный переход
- один символ исходного сообщения может заменяться одним символом нового кода или несколькими символами, а может быть и наоборот – несколько символов исходного сообщения заменяются одним символом в новом коде (китайские иероглифы обозначают целые слова и понятия)
- кодирование может быть *равномерное* и *неравномерное*; при равномерном кодировании все символы кодируются кодами равной длины; при неравномерном кодировании разные символы могут кодироваться кодами разной длины, это затрудняет декодирование
- закодированное сообщение можно однозначно декодировать с начала, если выполняется *условие Фано*: никакое кодовое слово не является началом другого кодового слова;
- закодированное сообщение можно однозначно декодировать с конца, если выполняется *обратное условие Фано*: никакое кодовое слово не является окончанием другого кодового слова;
- условие Фано – это достаточное, но не необходимое условие однозначного декодирования.

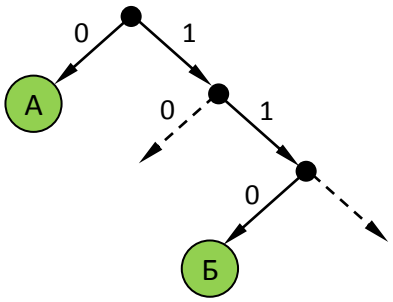
Пример задания

Р-12. *Для кодирования некоторой последовательности, состоящей из букв А, Б, В, Г, решили использовать неравномерный двоичный код, удовлетворяющий условию Фано. Для буквы А использовали кодовое слово 0, для буквы Б – кодовое слово 110. Какова наименьшая возможная суммарная длина всех четырёх кодовых слов?*

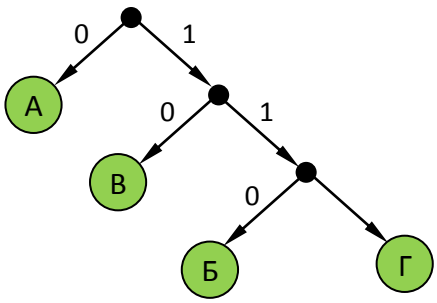
1) 7 2) 8 3) 9 4) 10

Решение:

1) условие Фано означает, что ни одно кодовое слово не совпадает с началом другого кодового слова; при этом в дереве кода все кодовые слова должны располагаться в листьях дерева, то есть в узлах, которые не имеют потомков;
2) построим дерево для заданных кодовых слов А – 0 и Б – 110:



3) штриховыми линиями отмечены две «пустые» ветви, на которые можно «прикрепить» листья для кодовых слов букв В (10) и Г (111)



4) таким образом, выбрав кодовые слова А – 0, Б – 110, В – 10, Г – 111, получаем суммарную длину кодовых слов 9 символов
5) Ответ: 3.

Ещё пример задания

Р-11. *По каналу связи передаются сообщения, содержащие только 5 букв А, И, К, О, Т. Для кодирования букв используется неравномерный двоичный код с такими кодовыми словами:*
А — 0, И — 00, К — 10, О — 110, Т — 111.
Среди приведённых ниже слов укажите такое, код которого можно декодировать только одним способом. Если таких слов несколько, укажите первое по алфавиту.

1) КАА 2) ИКОТА 3) КОТ 4) ни одно из сообщений не подходит

Решение:

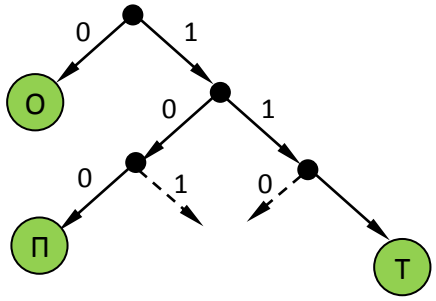
1) прежде всего заметим, что для заданного кода не выполняется ни прямое, ни обратное условие Фано; «виновата» в этом пара А – И: код буквы А совпадает как с началом, так и с окончанием кода буквы И; больше ни для одной пары кодовых слов прямое условие Фано не нарушено
2) это означает, что не все сообщения могут быть декодированы однозначно
3) теперь нужно понять, какие последовательности могут быть декодированы неоднозначно; в данном случае очевидно, что сообщения АА и И кодируются одинаково: 00, поэтому все слова, где есть АА или И, не могут быть декодированы однозначно
4) поэтому варианты 1 (КАА) и 2 (ИКОТА) отпадают
5) на всякий случай проверим вариант 3: КОТ = 10110111; первой буквой может быть только К (по-другому сочетание 10 получить нельзя), аналогично вторая буква – только О, а третья – только Т
6) Ответ: 3.

Ещё пример задания

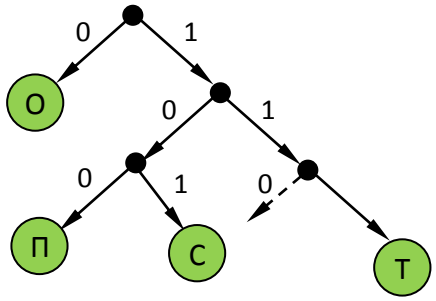
Р-10. *По каналу связи передаются сообщения, содержащие только 4 буквы П, О, С, Т; для передачи используется двоичный код, допускающий однозначное декодирование. Для букв Т, О, П используются такие кодовые слова: Т: 111, О: 0, П: 100.*
Укажите кратчайшее кодовое слово для буквы С, при котором код будет допускать однозначное декодирование. Если таких кодов несколько, укажите код с наименьшим числовым значением.

Решение:

1) условие Фано означает, что ни одно кодовое слово не совпадает с началом другого кодового слова; при этом в дереве кода все кодовые слова должны располагаться в листьях дерева, то есть в узлах, которые не имеют потомков;
2) построим дерево для заданных кодовых слов О – 0, Т – 111 и П – 100:



3) штриховыми линиями отмечены две «пустые» ветви, на которые можно «прикрепить» лист для кодового слова буквы С: 101 или 110; из них минимальное значение имеет код 101



4) таким образом, выбрав кодовые слова А – 0, Б – 110, В – 10, Г – 111, получаем суммарную длину кодовых слов 9 символов
5) Ответ: 101.

Ещё пример задания

Р-09. *Для кодирования некоторой последовательности, состоящей из букв А, Б, В, Г и Д, используется неравномерный двоичный код, позволяющий однозначно декодировать полученную двоичную последовательность. Вот этот код: А – 0; Б – 100; В – 1010; Г – 111; Д – 110. Требуется сократить для одной из букв длину кодового слова так, чтобы код по-прежнему можно было декодировать однозначно. Коды остальных букв меняться не должны.*
Каким из указанных способов это можно сделать?

1) для буквы В – 101
2) это невозможно
3) для буквы В – 010
4) для буквы Б – 10

Решение:

1) код однозначно декодируется, если выполняется условие Фано или обратное условие Фано; в данном случае «прямое» условие Фано выполняется: с кода буквы А (0) не начинается ни один другой код, оставшиеся короткие коды (Б, Г и Д) не совпадают с началом длинного кода буквы В; таким образом, при сокращении нужно сохранить выполнение условия Фано
2) вариант 3 не подходит, потому что новый код буквы В начинается с 0 (кода А), поэтому условие Фано нарушено
3) вариант 4 не подходит, потому что код буквы В начинается с 10 (нового кода б), поэтому условие Фано нарушено
4) вариант 1 подходит, условие Фано сохраняется (все трёхбитные коды различны, ни один не начинается с 0)
5) Ответ: 1.

Ещё пример задания

Р-08. *По каналу связи передаются сообщения, содержащие только 4 буквы: А, И, С, Т.*
В любом сообщении больше всего букв А, следующая по частоте буква – С, затем – И. Буква Т встречается реже, чем любая другая. Для передачи сообщений нужно использовать неравномерный двоичный код, допускающий однозначное декодирование; при этом сообщения должны быть как можно короче. Шифровальщик может использовать один из перечисленных ниже кодов. Какой код ему следует выбрать?

1) А – 0, И – 1, С – 00, Т – 11
2) С – 1, И – 0, А – 01, Т – 10
3) А – 1, И – 01, С – 001, Т – 000
4) С – 0, И – 11, А – 101, Т – 100

Решение:

1) сначала выберем коды, допускающие однозначное декодирование: это коды 3 и 4 (для них выполняется условие Фано), коды 1 и 2 не подходят
2) для того, чтобы длина сообщения была как можно короче, должно выполнять правило: «чем чаще встречается буква, тем короче её код»;
3) к сожалению, правило, приведённое выше, не совсем «хорошо» выполняется для кодов 3 и 4: в коде 3 длина кодового слова для буквы С больше, чем длина кодового слова буквы И (а хочется наоборот); для кода 4 длина кодового слова для буквы А – не самая маленькая из всех
4) сравним коды 3 и 4, предполагая, что в сообщении буква А встречается α раз, буква С – β раз, буква И – γ раз и буква Т – δ раз; причём по условию задачи $\alpha > \beta > \gamma > \delta$
5) при кодировании кодом 3 получаем сообщение длиной
   $L_3 = \alpha + 3\beta + 2\gamma + 3\delta$
6) при кодировании кодом 4 получаем сообщение длиной
   $L_4 = 3\alpha + \beta + 2\gamma + 3\delta$
7) находим разность: $L_4 - L_3 = (3\alpha + \beta + 2\gamma + 3\delta) - (\alpha + 3\beta + 2\gamma + 3\delta) = 2\alpha - 2\beta$
8) поскольку $\alpha > \beta$, получаем $L_4 - L_3 > 0$, то есть код 3 более экономичный
9) Ответ: 3.

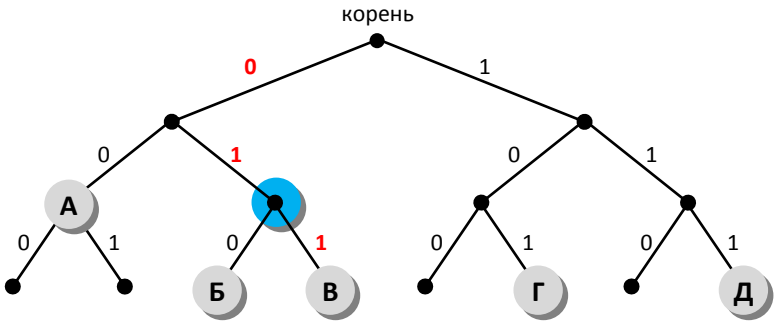
Ещё пример задания:

Р-06. *Для кодирования некоторой последовательности, состоящей из букв А, Б, В, Г и Д, используется неравномерный двоичный код, позволяющий однозначно декодировать полученную двоичную последовательность. Вот этот код: А–00, Б–010, В–011, Г–101, Д–111. Можно ли сократить для одной из букв длину кодового слова так, чтобы код по-прежнему можно было декодировать однозначно? Коды остальных букв меняться не должны. Выберите правильный вариант ответа.*

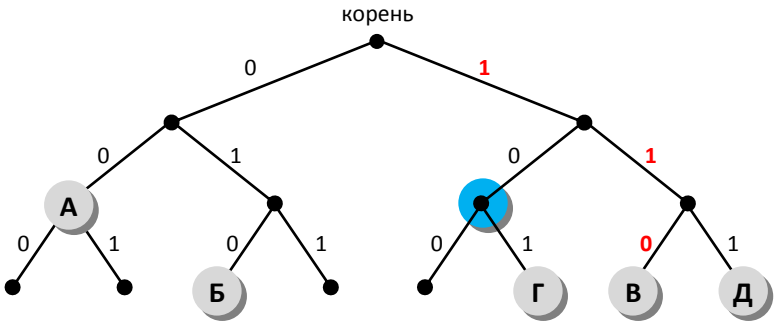
1) для буквы Б – 01
2) это невозможно
3) для буквы В – 01
4) для буквы Г – 01

Решение:

1) построим двоичное дерево, в котором от каждого узла отходит две ветки, соответствующие выбору следующей цифры кода – 0 или 1; разместим на этом дереве буквы А, Б, В, Г и Д так, чтобы их код получался как последовательность чисел на рёбрах, составляющих путь от корня до данной буквы (красным цветом выделен код буквы В – 011):



2) здесь однозначность декодирования получается за счёт того, что при движении от корня к любой букве в середине пути не встречается других букв (выполняется условие Фано);
3) теперь проверим варианты ответа: предлагается перенести одну из букв, Б, В или Г, в узел с кодом 01, выделенный синим цветом
4) видим, что при переносе любой из этих букв нарушится условие Фано; например, при переносе буквы Б в синий узел она оказывается на пути от корня до В, и т.д.; это значит, что предлагаемые варианты не позволяют выполнить прямое условие Фано
5) хочется уже выбрать вариант 2 («это невозможно»), но у нас есть еще обратное условие Фано, для которого тоже можно построить аналогичное дерево, в котором движение от корня к букве дает её код **с конца** (красным цветом выделен код буквы В – 011, записанный с конца):



видно, что обратное условие Фано также выполняется, потому что на пути от корня к любой букве нет других букв
6) в заданных вариантах ответа предлагается переместить букву Б, В или Г в синий узел; понятно, что Б или В туда перемещать нельзя – перемещённая буква отказывается на пути от корня к букве Г; а вот букву Г переместить можно, при этом обратное условие Фано сохранится
7) правильный ответ – 4.

Ещё пример задания:

Р-05. *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:

1) заметим, что для известной части кода выполняется условие Фано – никакое кодовое слово не является началом другого кодового слова
2) если Д = 00, такая кодовая цепочка совпадает с началом Б = 000 и В = 001, невозможно однозначно раскодировать цепочку 000000: это может быть ДДД или ББ; поэтому первый вариант не подходит
3) если Д = 01, такая кодовая цепочка совпадает с началом Г = 011, невозможно однозначно раскодировать цепочку 011: это может быть ДА или Г; поэтому второй вариант тоже не подходит
4) если Д = 11, условие Фано тоже нарушено: кодовое слово А = 1 совпадает с началом кода буквы Д, невозможно однозначно раскодировать цепочку 111: это может быть ДА или ААА; третий вариант не подходит
5) для четвертого варианта, Д = 010, условие Фано не нарушено;
6) правильный ответ – 4.

Возможные ловушки:

- условие Фано – это **достаточное**, но не необходимое условие однозначного декодирования, поэтому для уверенности полезно найти для всех «неправильных» вариантов контрпримеры: цепочки, для которых однозначное декодирование невозможно