**Тема 2.1. Тексты и кодирование**

***Равномерные и неравномерные коды***

Код называется **равномерным** (или кодом **постоянной длины**), если все его кодовые слова содержат одинаковое число букв (одинаковую длину слов). Соответственно, кодирование называется **равномерным**, если соответствующий ему код имеет постоянную длину. В настоящее время в информатике более употребительно равномерное кодирование, оно проще и более удобно. В компьютерах при кодировании информации в основном используются равномерные коды, соответствующие размерам компьютерных ячеек.

К равномерным кодам относится телеграфный код Бодо (Baudot code). Его можно считать двоичным равномерным алфавитным кодом. Первоначальный вариант этого кода разработал Эмиль Бодо в 1870 году для своего телеграфа. Код вводился прямо клавиатурой, состоящей из пяти клавиш, нажатие или ненажатие клавиши соответствовало передаче или непередаче одного бита в пятибитном коде. Например, буква **А** передавалась как **- - + - -** , что соответствовало нажатию средней клавиши. В двоичном коде это можно записать как **00100**. Таким образом, каждая буква записывалась пятью битами. Следовательно, кодом Бодо можно было передать **25=32** различных символа.

Другим интересным примером равномерного кода является код Трисиме, в котором знакам латинского алфавита ставятся в соответствие кодовые слова длины 3 над алфавитом из 3-х символов: {1, 2, 3}. Этот код представлен в следующей таблице:



Понятно, что код Трисиме не может кодировать более чем **33=27** символов.

Число букв в алфавите кода называется **основанием** кода, а длина кодовых слов равномерного кода называется **порядком** кода. Коды с основанием 2, как уже говорилось, называются **двоичными**, а с основанием 3 – **троичными**, и так далее. Так код Бодо имеет основание 2, а порядок 5, а у кода Трисиме и основание, и порядок равны 3.

Код называется **неравномерным** (или кодом **переменной длины**), если его кодовые слова имеют разное число букв (неодинаковую длину слов). Соответственно, кодирование называется **неравномерным,** если соответствующий ему код неравномерный.

Типичным примером неравномерного кода является телеграфный код, который принято называть азбукой Морзе. На следующей таблице представлен код азбуки Морзе для русского алфавита:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| A | • − | И | • • | P | • − • | Ш | − − − − | http://ok-t.ru/life-prog/baza2/4833556307348.files/image004.jpg*Сэмюэл Морзе (1791-1872)* |
| Б | − • • • | Й | • − − − | С | • • • | Щ | − − • − |
| В | • − − | К | − • − | Т | − | Ъ | • − − • − • |
| Г | − − • | Л | • − • • | У | • • − | Ь | − • • − |
| Д | − • • | М | − − | Ф | • • − • | Ы | − • − − |
| Е | • | H | − • | Х | • • • • | Э | • • − • • |
| Ж | • • • − | О | − − − | Ц | − • − • | Ю | • • − − |
| З | − − • • | П | • − − • | Ч | − − − • | Я | • − • − |

Как видим, азбука Морзе состоит из слов над алфавитом из двух символов: точка и тире. Но, строго говоря, этот код не является двоичным, так как он при кодировании слов предполагает еще один символ для разделения букв в слове (символ «пауза»). Без этого символа не было бы однозначности при декодировании текстов. Например, код из четырех тире можно было бы декодировать по-разному: или как код одной буквы Ш, или как код сочетаний из двух букв - ММ, ОТ или ТО. Разделяющий символ позволяет однозначно декодировать любую кодовую последовательность, полученную кодированием сообщений при помощи азбуки Морзе, но тогда код азбуки Морзе следует считать троичным, так как его алфавит содержит три символа.

Американский изобретатель телеграфа Сэмюель Морзе разработал этот код в 1838 году для передачи телеграфных сообщений в виде последовательности электрических сигналов, передаваемых от одного телеграфного аппарата по проводам к другому телеграфному аппарату. Этот код был придуман Морзе задолго до научных исследований относительной частоты появления различных букв в текстах, но, тем не менее, Морзе при составлении кода использовал принцип частоты букв. Буквам, используемым чаще, им присвоены короткие кодовые комбинации, редко используемым буквам – длинные. Морзе оценил относительную частоту букв английского языка подсчетом литер в ячейках типографской наборной машины. Наиболее часто используемой букве «Е» (в английском языке) он присвоил наиболее короткий код «точка». Следующей по количеству литер букве он присвоил код несколько большей длительности и так далее.

При составлении азбуки Морзе для букв русского алфавита учет относительной частоты букв не производился, и это повысило его избыточность. Расчеты избыточности кода Морзе на основании проведенных исследований частоты появления букв показали, что для букв английского алфавита она составляет 19%, для букв русского алфавита 22%.

Самым знаменитым телеграфным сообщением является сигнал бедствия "SOS" (Save Our Souls - спасите наши души). Вот как он выглядит: «• • • – – – • • •»

Преимущество у неравномерных кодов перед равномерными как раз и состоит в том, что сообщения можно передавать более экономным способом, так как часто передаваемые кодовые слова более короткие, а значит, кодовая последовательность может иметь меньшую длину, чем для равномерных кодов. Ниже это будет показано.

Но у неравномерных кодов есть серьезный недостаток по сравнению с равномерными кодами. У равномерных кодов кодовая последовательность всегда декодируется однозначно за счет того, что кодовые слова имеют одинаковую длину (кодовая последовательность легко делится на кодовые слова). Но не для всех неравномерных кодов достигается однозначность декодирования кодовых последовательностей. Мы уже видели это, пытаясь рассматривать азбуку Морзе как двоичный код.

Условие Фано

Кодирование символов обычно предполагает, что каждому символу всегда сопоставляется одинаковое количество битов (например, в кодовой таблице ASCII каждому символу сопоставляется один байт, хранящий порядковый номер того или иного символа в этой таблице). Такой способ кодирования прост и удобен, однако очевидно, что он является не самым оптимальным. Для значительной части символов используются не все биты отведенных под них байтов (часть старших битов — нулевые), а при наличии в тексте только части символов, предусмотренных в таблице ASCII (например, если текст содержит только прописные русские буквы), приходится все равно использовать 8-битный код.

Более компактным является неравномерный двоичный код (особенно если при его построении исходить из частоты встречаемости различных символов и присваивать наиболее часто используемым знакам самые короткие коды, как это сделано в методе Хаффмана). При этом количество битов, отводимых для кодирования символов, в целом зависит от количества используемых в конкретном случае различных символов (от мощности алфавита), а коды, соответствующие разным символам, могут иметь различную длину в битах.

Главное при таком кодировании — обеспечить возможность однозначного декодирования записанной с помощью этих кодов строки (поочередного, слева направо, выделения и распознавания из сплошной последовательности нулей и единиц кодов отдельных букв). Для этого коды символам необходимо назначать в соответствии с условиями Фано.

Прямое условие Фано. Неравномерный код может быть однозначно декодирован, если никакой из кодов не совпадает с началом (префиксом) какого-либо другого, более длинного кода.



Обратное условие Фано. Неравномерный код может быть однозначно декодирован, если никакой из кодов не совпадает с окончанием (постфиксом) какого-либо другого, более длинного кода.



Для однозначности декодирования последовательности кодов достаточно выполнения хотя бы одного из двух вышеуказанных условий Фано:

— при выполнении прямого условия Фано последовательность кодов однозначно декодируется с начала;

— при выполнении обратного условия Фано последовательность кодов однозначно декодируется с конца.

Выбрать, какое из двух правил Фано используется при решении конкретной задачи, можно, проанализировав коды в условии задачи (без учёта кода, проверяемого в вариантах ответа): если для исходных кодов выполняется прямое правило Фано, то его и нужно использовать при решении, и наоборот.

Вместе с тем нужно помнить, что правила Фано — это достаточное, но не необходимое условие однозначного декодирования: если не выполняется ни прямое, ни обратное правило Фано, конкретная двоичная последовательность может оказаться такой, что она декодируется однозначно (так как остальные возможные варианты до конца декодирования довести не удаётся). В подобном случае необходимо пытаться строить дерево декодирования в обоих направлениях.

Рекомендуется начинать решение задач такого типа с анализа выполнимости правил Фано для исходных кодов, указанных в условии задачи (т.е. без учета искомого кода в вариантах ответов). В зависимости от того, какое из двух правил Фано выполняется для исходных кодов, при дальнейшем решении задачи производится сравнение более короткого кода с началом (при выполнении прямого правила Фано) или с концом (при выполнении обратного правила Фано) более длинного кода.

Если для заданной последовательности кодов выполняется прямое правило Фано, то её декодирование необходимо вести с начала (слева направо).

Если для заданной последовательности кодов выполняется обратное правило Фано, то её декодирование необходимо вести с конца (справа налево).

При сравнении пары кодов удобно подписывать более короткий код под более длинным, выравнивая эти записи по левому краю — для прямого правила Фано либо по правому краю — для обратного правила Фано.

Задача 1.

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.

*Решение*

|  |  |  |
| --- | --- | --- |
| Проверяемый кодбуквы Д | Существующие коды букв А, Б, В и Г | Вывод |
| А | Б | В | Г |
| 1 | 000 | 001 | 011 |
| 00 | 001Совпадения нет | 00000Совпадение есть! | 00100Совпадение есть! | 01100Совпадения нет | Код 00 для буквы Д непригоден |
| 01 | 011Совпадения нет | 00001Совпадения нет | 00101Совпадения нет | 01101Совпадение есть! | Код 01 для буквы Д непригоден |
| 11 | 111Совпадение есть! | 00011Совпадения нет | 00111Совпадения нет | 01111Совпадения нет | Код 11 для буквы Д непригоден |
| 010 | 0101Совпадения нет | 000010Коды не равны | 001010Коды не равны | 011010Коды не равны | Код 010 для буквы Д пригоден |

В итоге, допустимым для буквы Д является код 010. Это единственный возможный вариант, поэтому условие о том, что кодовое слово должно иметь минимально возможную длину, в данном случае не требуется.

Ответ: код 010.

*Задача 2.*

Для кодирования последовательности символов, состоящей из букв К, И, Н, О, используется неравномерный двоичный код, удовлетворяющий условию Фано. При этом для буквы К использован код 0, а для буквы И — код 11. Требуется определить наименьшую возможную суммарную длину всех кодовых слов указанных букв.

*Решение*

Для проверки на соответствие кодов условию Фано нужно попарно сравнивать между собой коды по следующим правилам:

• когда длина обоих сравниваемых кодов совпадает, проверяется равенство этих кодов: если один код совпадает с другим, то такая пара кодов неправильна (не удовлетворяет условию Фано);

• когда длина сравниваемых кодов различна, более короткий код записывается под более длинным с выравниванием обоих кодов по левому краю: если все знаки более короткого кода совпадают с соответствующими знаками в начале более длинного кода, то такая пара кодов неправильна (не удовлетворяет условию Фано).

Опираясь на эти правила, подбираются коды для оставшихся букв Н и О. Начнём с буквы Н и будем перебирать возможные двоичные числа с возрастающей длиной (важно получить наименьшую возможную суммарную длину кодов!).

|  |  |  |  |
| --- | --- | --- | --- |
| Код буквы К | Код буквы И | Предполагаемый код буквы Н | Комментарий |
| 0 | 11 | 1 | Нельзя, так как совпадает с началом кода буквы И |
|  |  | 00 | Нельзя — код буквы К совпадает с его началом |
|  |  | 01 | Нельзя — код буквы К совпадает с его началом |
|  |  | 10 | Допустимый код (не совпадает с двузначным кодом буквы И, а код буквы К не совпадает с его началом) |

Итак, можно предположить, что первый код найден. Но посмотрим — удастся ли при этом найти код для оставшейся четвёртой буквы О. При этом можно сразу отбросить те коды, которые не подошли для буквы Н, — ведь код буквы О должен удовлетворять тем же требованиям при сравнении с кодами букв К и И.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Код буквы К | Код буквы И | Код буквы Н | Предполагаемый код буквы О | Комментарий |
| 0 | 11 | 10 | 11 | Нельзя — совпадает с кодом буквы И |
|  |  |  | 000, 001, 010, 011 | Нельзя — код буквы К совпадает с его началом |
|  |  |  | 100, 101 | Нельзя — код буквы Н совпадает с его началом |
|  |  |  | 110, 111 | Нельзя — код буквы Н совпадает с его началом |

Очевидно, и дальше с увеличением числа разрядов в предполагаемом коде буквы О будет сохраняться та же тенденция: ни один код не пригоден. Как же быть?

Причина этой прискорбной ситуации — в том, что, выбрав для буквы Н код 10, мы “закрыли” для себя возможность дальнейшего расширения нашей кодовой системы. Поэтому вместо кода 10 придётся выбрать для буквы Н более длинный код, — например, 100.

Теперь повторим попытку поиска кода для буквы О:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Код буквы К | Код буквы И | Код буквы Н | Предполагаемый код буквы О | Комментарий |
| 0 | 11 | 100 | 101 | Допустимый код (не совпадает с трёхзначным кодом буквы Н, а коды букв К и И не совпадают с его началом) |

Таким образом, решение найдено. Выпишем коды всех четырёх букв:

|  |  |  |  |
| --- | --- | --- | --- |
| К | И | Н | О |
| 0 | 11 | 100 | 101 |

Подсчитаем суммарную длину этих кодов (в знаках): 1 + 2 + 3 + 3 = 9.

Ответ: 9.

Задача 3.

При передаче информации используется равномерный двоичный код. Передаче подлежат сообщения, которые могут состоять только из четырёх букв: “И”, “Н”, “Ф”, “О”, при этом каждой букве соответствует отдельное кодовое слово.

Для используемого набора кодовых слов выполняется обязательное правило: любые два кодовых слова из этого набора должны различаться как минимум в трёх разрядах.

Для кодирования букв “И”, “Н”, “О” используются следующие 5-битовые кодовые слова:

“И”: 11110, “Н”: 10000, “О”: 01001. Про 5-битовый код для буквы “Ф” известно, что оно начинается с 0 и заканчивается 1.

Укажите кодовое слово для буквы “Ф”.

Решение

В задачах этого типа речь идёт об обычном двоичном коде, длина которого одинакова для всех кодируемых символов. Условия Фано в этом случае не используются, а для решения задачи достаточно обычных рассуждений.

1. По условию, любые два кодовых слова из используемого набора различаются как минимум в трёх разрядах. Запишем предполагаемый код буквы “Ф”, заменяя неизвестные пока биты звездочками: “Ф”: 0\*\*\*1.

2. Сравним этот код с известными кодами других букв:

“И”: 11110

“Н”: 10000

“О”: 01001

“Ф”: 0\*\*\*1

Видим, что начальный и конечный биты кода буквы “Ф” не совпадают с соответствующими битами кодов букв “И” и “Н”, но совпадают (причём оба!) с начальным и конечным битами для буквы “О”.

3. По условию, нужно, чтобы коды “О” и “Ф” различались как минимум в трёх позициях. А такое различие нам могут дать только три внутренних бита.

Тогда для “Ф” вместо звёздочек надо выбрать биты, противоположные соответствующим битам для “О”:

“О”: 01001

“Ф”: 00111

3. Для кодов “Ф” и “О” требуемое условие соблюдено. Проверим его соблюдение для кодов остальных букв:

“И”: 11110 “Н”: 10000

“Ф”: 00111 “Ф”: 00111

3 различия 4 различия

Таким образом, условие соблюдено и для этих кодов.

Ответ: 10110.