

<https://habr.com/ru/post/231177/>

1. Что означает алгоритм сжатия с потерями и без потерь. Приведите примеры где они применяются.
2. Выбрать три алгоритма кодирования использующиеся в архиваторах 7zip, winzip, winrar и т.п. Описать суть метода кодирования. Достоинства. Недостатки.
3. Переписать задачи и конспект:

Пример18: Провести эффективное кодирование ансамбля из восьми знаков ($m=8$), используя метод Шеннона - Фано.

Решение: При обычном (не учитывающем статистических характеристик) двоичном кодировании с использованием $k=2$ знаков при построении равномерного кода количество элементов в кодовой последовательности будет $q \geq \log_k m = \log_2 8 = 3$, т.е. для представления каждого знака использованного алфавита потребуется три двоичных символа.

Метод Шеннона - Фано позволяет построить кодовые комбинации, в которых знаки исходного ансамбля, имеющие наибольшую вероятность, кодируются наиболее короткими кодовыми последовательностями. Таким образом, устраняется избыточность обычного двоичного кодирования, информационные возможности которого используются не полностью.

Таблица 3.2.

Знаки (буквы) x_i	Вероятность P_i	Кодовые комбинации						
		номер разбиения						
		1	2	3	4	5	6	7
x1	1/2	1						
x2	1/4	0	1					
x3	1/8	0	0	1				
x4	1/16	0	0	0	1			
x5	1/32	0	0	0	0	1		
x6	1/64	0	0	0	0	0	1	
x7	1/128	0	0	0	0	0	0	1

x8	1/128	0	0	0	0	0	0	0
----	-------	---	---	---	---	---	---	---

Так как вероятности знаков представляют собой отрицательные целочисленные степени двойки, то избыточность при кодировании устранена полностью.

Среднее число символов на знак в этом случае точно равно энтропии. В общем случае для алфавита из восьми знаков среднее число символов на знак будет меньше трех, но больше энтропии алфавита. Вычислим энтропию алфавита:

$$H = - \sum_{i=1}^{m=8} P(x_i) \log P(x_i) = 1 \frac{63}{64}$$

Вычислим среднее число символов на знак:

$$q_{cp} = \sum_{i=1}^{m=8} P(x_i) q(x_i) = 1 \frac{63}{64},$$

где $q(x_i)$ - число символов в кодовой комбинации, соответствующей знаку x_i .

Пример19: Определить среднюю длину кодовой комбинации при эффективном кодировании по методу Шеннона - Фано ансамбля - из восьми знаков и энтропию алфавита.

Таблица 3.3.

Знаки (буквы) x_i	Вероятность P_i	Кодовые комбинации				
		номер разбиения				
		1	2	3	4	5
x1	0,22	1	1			
x2	0,20	1	0	1		
x3	0,16	1	0	0		
x4	0,16	0	1			
x5	0,10	0	0	1		
x6	0,10	0	0	0	1	
x7	0,04	0	0	0	0	1
x8	0,02	0	0	0	0	0

Решение: 1. Средняя длина кодовых комбинаций

$$q_{cp} = \sum_{i=1}^{m=8} P_i q_i = 2,84$$

2. Энтропия алфавита

$$H = - \sum_{i=1}^{m=8} P_i \log P_i = 2,76$$

При кодировании по методу Шеннона - Фано некоторая избыточность в последовательностях символов, как правило, остается ($q_{cp} > H$).

Эту избыточность можно устранить, если перейти к кодированию достаточно большими блоками.

4. Кодирование информации для канала с помехами

Ошибка в кодовой комбинации появляется при ее передаче по каналу связи вследствие замены одних элементов другими под воздействием помех. Например, 2-кратная ошибка возникает при замене (искажении) двух элементов. Например, если кодовая комбинация 0110111 принята как 0100110, то имеет место двукратная ошибка.

Теория помехоустойчивого кодирования базируется на результатах исследований, проведенных Шенноном и сформулированных в виде теоремы:

1. При любой производительности источника сообщений, меньшей, чем пропускная способность канала, существует такой способ кодирования, который позволяет обеспечить передачу всей информации, создаваемой источником сообщений, со сколь угодно малой вероятностью ошибки.

2. Не существует способа кодирования, позволяющего вести передачу информации со сколь угодно малой вероятностью ошибки, если производительность источника сообщений больше пропускной способности канала.

Из теоремы следует, что помехи в канале не накладывают ограничений на точность передачи. Ограничение накладывается только на скорость передачи, при которой может быть достигнута сколь угодно высокая точность передачи.

Теорема не затрагивает вопроса о путях построения кодов, обеспечивающих идеальную передачу информации, но, обосновав принципиальную возможность такого кодирования, позволяет вести разработку конкретных кодов.

При любой конечной скорости передачи информации вплоть до пропускной способности канала, сколь угодно малая вероятность ошибки достигается лишь при безграничном увеличении длительности кодируемых последовательностей знаков. Таким образом, *безошибочная передача при наличии помех возможна лишь теоретически.*

Обеспечение передачи информации с весьма малой вероятностью ошибки и достаточно высокой эффективностью возможно при кодировании чрезвычайно длинными последовательностями знаков.

На практике точность передачи информации и эффективность каналов связи ограничивается двумя факторами:

- 1) размером и стоимостью аппаратуры кодирования/декодирования;
- 2) временем задержки передаваемого сообщения.

4.1 Разновидности помехоустойчивых кодов

Коды, которые обеспечивают возможность обнаружения и исправления ошибки, называют помехоустойчивыми.

Эти коды используют для:

- 1) исправления ошибок – корректирующие коды;
- 2) обнаружения ошибок.

Корректирующие коды основаны на введении избыточности.

У подавляющего большинства помехоустойчивых кодов помехоустойчивость обеспечивается их алгебраической структурой. Поэтому их называют алгебраическими кодами.

Алгебраические коды подразделяются на два класса:

- 1) блочные;
- 2) непрерывные.

В случае блочных кодов процедура кодирования заключается в сопоставлении каждой букве сообщения (или последовательности из k символов, соответствующей этой букве) блока из n символов. В операциях по преобразованию принимают участие только указанные k символов, и выходная последовательность не зависит от других символов в передаваемом сообщении.

Блочный код называют равномерным, если n остается постоянным для всех букв сообщения.

Различают *разделимые и неразделимые блочные коды*. При кодировании *разделимыми кодами* выходные последовательности состоят из символов, роль которых может быть отчетливо разграничена. Это информационные символы, совпадающие с символами последовательности, поступающей на вход кодера канала, и избыточные (проверочные) символы, вводимые в исходную последовательность кодером канала и служащие для обнаружения и исправления ошибок.

При кодировании *неразделимыми кодами* разделить символы входной последовательности на информационные и проверочные невозможно.

Непрерывными (древовидными) называют такие коды, в которых введение избыточных символов в кодируемую последовательность информационных символов осуществляется непрерывно, без разделения ее на независимые блоки. Непрерывные коды также могут быть разделимыми и неразделимыми.