ОБРАЗЕЦ РЕШЕНИЯ ЗАДАЧИ № 3.

Код Шеннона-Фано

Под энтропией (степенью неопределённости), связанной с одним символом, подразумевается взвешенный логарифм вероятности встречи этого символа в сообщении. Для конечного множества событий (букв алфавита) X = {x_1, x_2, …, x_n}, наступающих с вероятностью p_1, p_2, …, p_n ($\Sigma p_i=1$), Энтропия наступления одного события (получения одного символа) равна:

$$H_i = -p_i \ln p_i$$

Для независимых событий, энтропия последовательного наступления нескольких событий равна сумме энтропии этих событий.

Для случая отсутствия статистической взаимосвязи между буквами конструктивные методы построения эффективных кодов были даны впервые Шенноном и Фано. Их методики существенно не отличаются, и поэтому соответствующий код получил название кода Шеннона-Фано.

Рассмотрим алфавит из 8 букв. Ясно, что при обычном кодировании (не учитывающем статистических характеристик) для представления каждой буквы требуется 3 символа.

Наибольший эффект сжатия получается в случае, когда вероятности представляют собой отрицательные целочисленные степени двойки. Среднее число символов на букву в этом случае точно равно энтропии. В более общем случае для алфавита из 8 букв среднее число символов на букву будет меньше 3, но больше энтропии алфавита H(A).

Код строится следующим образом: буквы алфавита сообщений выписываются в таблицу в порядке убывания вероятностей. Затем разделяем их на две группы так, чтобы суммы вероятностей в каждой из групп были по возможности одинаковы. Всем буквам одной половины в качестве первого символа записывается 1, а всем другой - 0. Каждая из полученных групп, в свою очередь, разбивается на две подгруппы с одинаковыми суммарными вероятностями и т.д. Процесс повторяется до тех пор, пока в каждой подгруппе не останется по одной букве.

Рассчитаем среднюю длину полученных кодовых слов по формуле:

$$l_{cp} = \sum_{i=1}^{5} l_i \cdot P(x_i)$$

.

Найдем также минимальную среднюю длину кодового слова по формуле:

$$l_{cp.\min} = H(x) = -\sum_{i=1}^{5} P(x_i) \cdot \log_2 P(x_i)$$

.

Сравнивая эти два значения, можно заметить, что некоторая избыточность в последовательностях символов осталась. Из теории Шеннона следует, что эту избыточность можно устранить, если перейти к кодированию достаточно большими блоками.

Теоретический минимум H(A) может быть достигнут при кодировании блоков, включающих бесконечное число букв.

Рассмотренная методика Шеннона-Фано не всегда приводит к однозначному построению кода, так как, разбивая на подгруппы иначе, код может оказаться не самым лучшим.

Расшифровка текста производится однозначно. В связи с этим данный код называется префиксным. Никакое кодовое слово префиксного кода не является началом другого кодового слова.

Примеры решения задач

110) Простейший дискретный источник (n=5) описывается схемой:

$$X = \begin{Bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 \\ P(x_1) & P(x_2) & P(x_3) & P(x_4) & P(x_5) \end{Bmatrix}$$

Даны 5 случайных чисел: 3, 47, 43, 73, 86 для расчета вероятностей этих сообщений. Требуется найти среднюю длину кодового слова этого источника и укрупненного методом Шеннона-Фано.

Решение:

1). Рассчитаем вероятности сообщений:

$$S=3+47+43+73+86=252;$$

$$P(x_1)=\frac{3}{252}=0{,}012;\ P(x_2)=\frac{47}{252}=0{,}187;\ P(x_3)=\frac{43}{252}=0{,}171;\ P(x_4)=\frac{73}{252}=0{,}289;$$

$$P(x_5)=\frac{86}{252}=0{,}341.$$

2). Кодируем сообщения источника кодом Шеннона-Фано.

а). Выписываем все сообщения в таблицу по степени убывания вероятности:

x_5	x_4	x_2	x_3	x_1
0,341	0,289	0,187	0,171	0,012

б). Разобьём сообщения на две группы равной (насколько это возможно) вероятности. Первой группе присвоим символ 1, второй -0. Затем каждая из групп вновь делится на две подгруппы равной вероятности, которым тоже присваиваются символы 1 и 0. В результате многократного повторения этой процедуры получим таблицу кодовых слов.

x_5	x_4	x_2	x_3	x_1
1	*1*	*0*	*0*	*0*
1	*0*	*1*	*0*	*0*
			1	*0*

в). Рассчитаем среднюю длину кодового слова по формуле

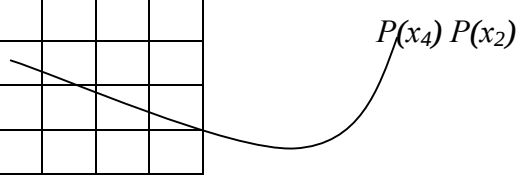
$$l_{ср}=\sum_{i=1}^{5} l_i \cdot P(x_i).$$

Вероятности сообщений	*0,341*	*0,289*	*0,187*	*0,171*	*0,012*
Длина кодовых слов l_i	*2*	*2*	*2*	*3*	*3*
$P(x_i)*l_i$	*0,682*	*0,578*	*0,374*	*0,513*	*0,036*

$$l_{ср.}=0{,}682+0{,}578+0{,}374+0{,}513+0{,}012=2{,}183.$$

3). Укрупним сообщения источника, переходя к блокам из двух сообщений по схеме:

	x_5	x_4	x_2	x_3	x_1
x_5					
x_4					
x_2					
x_3					
x_1					



Получаем следующую матрицу:

	0,341	0,289	0,187	0,171	0,012

0,341	0,116281	0,098549	0,063767	0,058311	0,004092
0,289	0,098549	0,083521	0,054043	0,049419	0,003468
0,187	0,063767	0,054043	0,034969	0,031977	0,002244
0,171	0,058311	0,049419	0,031977	0,029241	0,002052
0,012	0,004092	0,003468	0,002244	0,002052	0,000144

4). Кодируем укрупненные сообщения источника кодом Шеннона-Фано и рассчитаем среднюю длину кодового слова по формуле

$$l_{cp} = \frac{1}{2}\sum_{i=1}^{25} l_i \cdot P_i \ .$$

.

0,11628	0,09855	0,09855	0,08352	0,06377	0,06377	0,05831	0,05831	0,05404	0,05404	0,04942	0,04942	0,03497	0,03198	0,03198	0,02924	0,00409	0,00409	0,00347	0,00347	0,00224	0,00224	0,00205	0,00205	0,00014
1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	1	0	0	1	1	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
	1	0		1	0	1	0	1	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0
											1	0	1	0	1	1	0	0	0	0	0	0	0	0
															1	0	1	1	1	0	0	0	0	0
																	1	0	0	1	1	0	0	0
																		1	0	1	0	1	0	0
																							1	0

P_i	0,11628	0,09855	0,09855	0,08352	0,06377	0,06377	0,05831	0,05831	0,05404	0,05404	0,04942	0,04942	0,03497	0,03198	0,03198	0,02924	0,00409	0,00409	0,00347	0,00347	0,00224	0,00224	0,00205	0,00205
l_i	3	4	4	3	4	4	4	4	4	4	4	5	5	5	5	6	6	7	8	8	8	8	8	9
P_i*l_i	0,34884	0,3942	0,3942	0,25056	0,25507	0,25507	0,23324	0,23324	0,21617	0,21617	0,19768	0,2471	0,17485	0,15989	0,15989	0,17545	0,02455	0,02864	0,02774	0,02774	0,01795	0,01795	0,01642	0,01847

$$l_{cp.}=4{,}0924/2 \approx 2{,}046$$

5). Рассчитаем $l_{cp.\min} = H(x) = -\sum_{i=1}^{5} P(x_i)\cdot \log_2 P(x_i)$.

$P(x_i)$	0,341	0,289	0,187	0,171	0,012
$Log_2 P(x_i)$	-1,5521564	-1,790859	-2,4188898	-2,5479318	-6,3808218
$P(x_i)\ Log_2 P(x_i)$	-0,5292853	-0,517558	-0,4523324	-0,4356963	-0,0765699

$$l_{cp.min}=-(-2{,}0114) \approx 2{,}011.$$

Видим, что при укрупнении источника средняя длина ближе к минимальной, следовательно, можно сделать вывод, что укрупнение источника уменьшает среднюю длину кодового слова, устремляя её к H(x).