КОДИРОВАНИЕ

Авторы: Ю. В. Прохоров

КОДИРОВАНИЕ, операция отождествления символов или групп символов одного *кода* с символами или группами символов другого кода. Необходимость К. возникает прежде всего из потребности приспособить форму сообщения к данному *каналу связи* или к.-л. другому устройству, предназначенному для преобразования или хранения *информации*. Так, с помощью телеграфных кодов сообщения, представленные в виде последовательности букв и цифр, преобразуются в определённые комбинации точек и тире. Др. целью К. является защита сообщений от несанкционированного доступа (см. *Криптография*).

Цель К. в *информации теории* состоит в уменьшении т. н. *избыточности сообщений* и влияния помех, искажающих сообщения при передаче по каналам связи (см. *Шеннона теорема*). Поэтому выбор нового кода стремятся согласовать со статистич. свойствами источника сообщений. В какой-то степени это согласование имеется уже в коде Морзе, в котором чаще встречающиеся буквы обозначаются более короткими комбинациями точек и тире.

Приёмы, применяемые в теории информации для достижения указанного согласования, можно пояснить на примере построения экономных двоичных кодов. Пусть канал связи может передавать только символы 0 и 1, затрачивая на каждый одно и то же время
t. Для уменьшения времени передачи (или для увеличения её скорости) целесообразно до передачи кодировать сообщения таким образом, чтобы средняя длина
L кодового обозначения была наименьшей. Пусть
$x_1, x_2, \ldots, x_n$ обозначают возможные сообщения некоторого источника, а
$p_1, p_2, \ldots, p_n$ – соответствующие им вероятности. Тогда, как устанавливается в теории

информации, при любом способе K.

$$L \geq H,$$

где

$H = \sum_{i=1}^{n} p_i \log_2(1/p_i)$ – *энтропия* источника. Равенство в формуле

(1) может не достигаться. Однако при любых

$p_1, p_2, \ldots, p_n$ существует метод K. (метод Шеннона – Фэно), для которого

$$L \leq H + 1.$$

Метод состоит в том, что сообщения располагаются в порядке убывания вероятностей и полученный ряд делится на две части с суммарными вероятностями, по возможности близкими друг к другу. В качестве первого двоичного знака принимают 0 в 1-й части и 1 – во 2-й. Таким же образом делят пополам каждую из частей и выбирают второй двоичный знак и т. д., пока не придут к частям, содержащим только по одному сообщению.

Пример 1. Пусть

$n = 4$ и

$p_1 = {}^{9}/_{16}$,

$p_2 = p_3 = {}^{3}/_{16}$,

$p_4 = {}^{1}/_{16}$. Применение метода иллюстрируется таблицей

x_i	p_i	Кодовое обозначение
x_1	${}^{9}/_{16}$	0
x_2	${}^{3}/_{16}$	1 0
x_3	${}^{3}/_{16}$	1 1 0
x_4	${}^{1}/_{16}$	1 1 1

В данном случае

$$L = 1 \cdot \frac{9}{16} + 2 \cdot \frac{3}{16} + 3 \cdot \frac{3}{16} + 3 \cdot \frac{1}{16} = \frac{27}{16} = 1,688,$$

причём никакой другой код не даёт для

L меньшего значения. В то же время

$H = 1,623$. Всё сказанное применимо и к случаю, когда алфавит нового кода содержит не две, как предполагалось выше, а

$m > 2$ букв. При этом величина

H в формулах

(1) и

(2) должна быть заменена величиной

$H/\log_2 m$.

Задача о сжатии записи сообщений в данном алфавите (т. е. задача об уменьшении избыточности) может быть решена на основе метода Шеннона – Фэно. Если сообщения представлены последовательностями длины

N букв из

m-буквенного алфавита, то их средняя длина

L_N после К. всегда удовлетворяет неравенству

$$L_N \geq NH/\log_2 m,$$

где

H – энтропия источника на букву. Кроме того, при сколь угодно малом

$\varepsilon > 0$ можно добиться выполнения при всех достаточно больших

N неравенства

$$L_N < N(H/\log_2 m + \varepsilon).$$

С этой целью используют К. блоками: по данному

ε выбирают достаточно большое натуральное число

s и делят каждое сообщение на равные части – блоки, содержащие по

s букв. Затем эти блоки кодируют методом Шеннона – Фэно в тот же алфавит. Тогда при достаточно больших

N будет выполнено неравенство

(3). Справедливость этого утверждения легче всего понять, рассматривая случай, когда источником является последовательность независимых символов 0 и 1, появляющихся соответственно с вероятностями

p и

q,

$0 < p < 1$. Энтропия на блок равна

s-кратной энтропии на одну букву, т. е. равна

$sH = s(p\log_2 1/p + q\log_2 1/q)$. Кодовое обозначение блока требует в среднем не более $sH + 1$ двоичных знаков. Поэтому для сообщения, состоящего из

N букв,

$$L_N \leq (1 + N/s)(sH + 1) = N(H + 1/s)(1 + s/H),$$

что при достаточно больших

s и

N/s приводит к неравенству

(3). При таком К. энтропия на букву приближается к своему макс. значению – единице, а избыточность – к нулю.

Пример 2. Пусть источником сообщений является последовательность независимых знаков 0 и 1, в которой вероятность появления нуля равна

$p = {}^3/_4$, а единицы –

$q = {}^1/_4$. Здесь энтропия

H на букву равна 0,811, а избыточность – 0,189. Наименьшие блоки ($s = 2$), т. е. 00, 01, 10, 11, имеют соответственно вероятности

$p^2 = {}^9/_{16}$,

$pq = {}^3/_{16}$,

$qp = {}^3/_{16}$,

$q^2 = {}^1/_{16}$. Применение метода Шеннона – Фэно приводит к следующему правилу К.:

00→0, 01→10, 10→110, 11→111. При этом, напр., сообщение 00111000… примет вид 01111100… На каждую букву сообщения в прежней форме приходится в среднем

27/32=0,844 буквы в новой форме (при нижней границе коэф. сжатия, равной H=0,811). Энтропия на букву в новой последовательности равна 0,811/0,844=0,961, а избыточность равна 0,039.