Доклады Академии наук СССР
1970. Том 194, № 5

УДК 519.54 **МАТЕМАТИКА**

В. С. ПРОСКУРОВ

ТЕОРЕМА О КОДИРОВАНИИ ВЫСКАЗЫВАНИЙ В ПРОИЗВОЛЬНЫХ ЯЗЫКАХ

(Представлено академиком И. М. Виноградовым 26 III 1970)

Проблема кодирования [1] содержательных или формальных текстов является одной из проблем обработки информации. Основной трудностью является нахождение эффективного взаимно однозначного кодирования. В работе [2] рассмотрены различные правила кодирования слов некоторого словаря, состоящего из N слов, основное из которых — операция свертывания ∇_k кодов слов, содержащих l_i букв, до кодов, содержащих k букв ($k \leqslant l_i$, $1 \leqslant i \leqslant N$). В теореме 2 работы [2] доказывается, что свертывание ∇_k — неоднозначное кодирование с вероятностью нарушения однозначности, стремящейся к нулю при $k \to l = \max\limits_{1 \leqslant i \leqslant N} \{l_i\}$. Причем кодирование становится однозначным лишь при $k = l = \max\limits_{1 \leqslant i \leqslant N} \{l_i\}$. Свертывание ∇_k не позволяет по коду длины k определить исходное слово. При распространении теоремы 2 на высказывания величина k соответственно увеличится.

Ниже доказывается теорема о том, что существует взаимно однозначное кодирование, позволяющее любое высказывание представить кодом любой наперед заданной длины (например, $k = 1$), по которому можно восстановить исходное высказывание.

Введем необходимые определения и обозначения.

Алфавит $\mathfrak{A}$ есть строго упорядоченная последовательность n попарно различных символов с порядковыми номерами $0, 1, 2, \ldots, n-1$:

$$\mathfrak{A} = a_0 a_1 a_2 \ldots a_{n-1} = (a_i,\ i = 0, 1, 2, \ldots, n-1). \tag{1}$$

Символ a_0 будем называть пробелом.

Всякая совокупность букв из алфавита $\mathfrak{A}$, не содержащая ни одного пробела, называется словом в алфавите $\mathfrak{A}$:

$$\sigma_l(\mathfrak{A}, i_1, i_2, \ldots, i_k, \ldots, i_l) = a_{i_1} a_{i_2} \ldots a_{i_k} \ldots a_{i_l}. \tag{2}$$

В дальнейшем всегда будем иметь в виду, что $1 \leqslant i_k \leqslant n-1$ для всех $k = 1, 2, \ldots, l$ при $l \geqslant 1$.

Число букв, входящих в данное слово, будем называть длиной слова в алфавите $\mathfrak{A}$ и обозначать через l.

По определению при $l = 0$ будем считать $k = 0$, $i_k = 0$, $\sigma_0(\mathfrak{A}, 0) = a_0$ — пустое слово в алфавите $\mathfrak{A}$.

Словарем $\Sigma_N(\mathfrak{A})$ в алфавите $\mathfrak{A}$ назовем подмножество $\{\sigma_{l_r}(\mathfrak{A}, i_1, i_2, \ldots, i_k, \ldots, i_{l_r})\}$ множества $\Sigma(\mathfrak{A})$ всех слов в алфавите $\mathfrak{A}$, состоящее из N слов, где $1 \leqslant r \leqslant N$:

$$\Sigma_N(\mathfrak{A}) = \{\sigma_{l_r}(\mathfrak{A}, i_1, i_2, \ldots i_k, \ldots, i_{l_r})\} \subset \Sigma(\mathfrak{A}), \tag{3}$$

где $r = 1, 2, \ldots, N$.

Аналогично введем:

$$\mathfrak{B} = b_0 b_1 b_2 \ldots b_{m-1} = (b_i,\ i = 0, 1, 2, \ldots, m-1), \tag{4}$$

b_0 — пробел в алфавите $\mathfrak{B}$,

$$\sigma_l(\mathfrak{B}, i_1', i_2', \ldots, i_k', \ldots, i_l') = b_{i_1'} b_{i_2'} \ldots b_{i_k'} \ldots b_{i_l'} \tag{5}$$

слово в алфавите $\mathfrak{B}$, $1 \leqslant i \leqslant m-1$ для всех $k=1, 2, \ldots, l$ при $l \geqslant 1$, где l — длина слова в алфавите $\mathfrak{B}$.

По определению при $l=0$ будем считать $k=0$, $i_k'=0$ $\sigma_0(\mathfrak{B}, 0)=b_0$ — пустое слово в алфавите $\mathfrak{B}$.

$$\Sigma_N(\mathfrak{B}) = \{\sigma_{l_r}(\mathfrak{B}, i_1', i_2', \ldots, i_k', \ldots, i_{l_r}')\} \subset \Sigma(\mathfrak{B}), \tag{6}$$

где $r=1, 2, \ldots, N$; $\Sigma_N(\mathfrak{B})$ — словарь в алфавите $\mathfrak{B}$, содержащий N слов; $\Sigma(\mathfrak{B})$ — множество всевозможных слов в алфавите $\mathfrak{B}$.

Будем называть $K = (K'(\mathfrak{A}, \mathfrak{B}), K''(\mathfrak{B}, \mathfrak{A}))$ взаимно однозначным кодированием слов (или просто кодированием) в алфавите $\mathfrak{A}$ словами в алфавите $\mathfrak{B}$ и обратно, если

$$K'(\mathfrak{A}, \mathfrak{B})\ (\sigma_{l_1}(\mathfrak{A},\ i_1,\ i_2, \ldots, i_k, \ldots, i_{l_1})) = \sigma_{l_2}(\mathfrak{B},\ i_1',\ i_2', \ldots, i_k', \ldots, i_{l_2}'), \tag{7}$$

$$K''(\mathfrak{B}, \mathfrak{A})\ (\sigma_{l_2}(\mathfrak{B},\ i_1',\ i_2', \ldots,\ i_k', \ldots,\ i_{l_2}')) = \sigma_{l_1}(\mathfrak{A},\ i_1,\ i_2, \ldots, i_k, \ldots, i_{l_1}), \tag{8}$$

где $K'(\mathfrak{A}, \mathfrak{B})$ — однозначное преобразование (кодирование) слова в алфавите $\mathfrak{A}$ в слово в алфавите $\mathfrak{B}$ (допускается случай $\mathfrak{A}=\mathfrak{B}$) для любого слова из $\Sigma(\mathfrak{A})$; $K''(\mathfrak{B}, \mathfrak{A})$ — однозначное преобразование (кодирование) слова в алфавите $\mathfrak{B}$ в слово в алфавите $\mathfrak{A}$ (допускается случай $\mathfrak{B}=\mathfrak{A}$) для любого слова из $\Sigma(\mathfrak{B})$.

Согласно определению кодирования K устанавливает взаимно однозначное соответствие между словами (2) из $\Sigma(\mathfrak{A})$ и (6) из $\Sigma(\mathfrak{B})$

$$\sigma_l(\mathfrak{A}, i_1, i_2, \ldots, i_k, \ldots, i_l) \overset{K}{\rightleftarrows} \sigma_{l'}(\mathfrak{B}, i_1', i_2', \ldots, i_k', \ldots, i_{l'}'). \tag{9}$$

Следовательно, зная $\Sigma_N(\mathfrak{A})$, с помощью кодирования $K'(\mathfrak{A}, \mathfrak{B})$ можно получить $\Sigma_N(\mathfrak{B})$ и обратно — зная $\Sigma_N(\mathfrak{B})$, с помощью кодирования $K''(\mathfrak{B}, \mathfrak{A})$ можно получить $\Sigma_N(\mathfrak{A})$.

Теорема 1. *Для любого словаря $\Sigma_N(\mathfrak{A}) \subset \Sigma(\mathfrak{A})$ существует взаимно однозначное кодирование K и алфавит $\mathfrak{B}$ такие, что для каждого $\sigma_{l_r}(\mathfrak{A},\ i_1,\ i_2, \ldots, i_k, \ldots,\ i_{l_r}) \in \Sigma_N(\mathfrak{A})$, $1 \leqslant r \leqslant N$, будут справедливы соотношения*:

$$K'(\mathfrak{A}, \mathfrak{B})(\sigma_{l_r}(\mathfrak{A}, i_1, i_2, \ldots, i_k, \ldots, i_{l_r})) = \sigma_l(\mathfrak{B}, i_1', i_2', \ldots, i_k', \ldots, i_l'),$$

$$K''(\mathfrak{B}, \mathfrak{A})(\sigma_l(\mathfrak{B}, i_1', i_2', \ldots, i_k', \ldots, i_l')) = \sigma_{l_r}(\mathfrak{A}, i_1, i_2, \ldots, i_k, \ldots, i_{l_r}),$$

причем l может быть любым наперед заданным целым положительным числом с условием $l > l_r$ или $l \leqslant l_r, l_{r-1}, l_{r-2}, \ldots, 1$.

Доказательство. Пусть задан алфавит (1). Возьмем позиционную систему счисления с основанием n и цифрами

$$0, 1, 2, \ldots, n-1 = (k, k=0, 1, 2, \ldots, n-1). \tag{10}$$

Установим взаимно однозначное соответствие между символами алфавита $\mathfrak{A}$ и цифрами системы счисления (10) так, чтобы символу $a_k \in \mathfrak{A}$, стоящему в последовательности (1) на k-м месте, соответствовала цифра k из последовательности (10), стоящая на k-м месте, и обратно. Обозначим правило, устанавливающее это соответствие, через F^n:

$$a_k \overset{F^n}{\rightleftarrows} k,\ 0 \leqslant k \leqslant n-1. \tag{11}$$

Возьмем слово $\sigma_{l_r}(\mathfrak{A},\ i_1,\ i_2, \ldots,\ i_k, \ldots,\ i_{l_r}) \in \Sigma_N(\mathfrak{A})$. В силу правила F^n (11) слову $\sigma_{l_r}(\mathfrak{A}, i_1, i_2, \ldots, i_k, \ldots, i_{l_r})$ соответствует одно и только одно целое положительное число

$$i_1 i_2 \ldots i_k \ldots i_{l_r}(n) = i_1 n^{l_r-1} + i_2 n^{l_r-2} + \ldots + i_{l_r} = \sum_{k=1}^{l_r} i_k n^{l_r-k}$$

в позиционной системе счисления с основанием n и обратно:

$$F^n(\sigma_{l_r}(\mathfrak{A}, i_1, i_2, \ldots, i_k, \ldots, i_{l_r})) = i_1 i_2 \ldots i_k \ldots i_{l_r}(n), \tag{12}$$

$$F^n(i_1 i_2 \ldots i_k \ldots i_{l_r}(n)) = \sigma_{l_r}(\mathfrak{A}, i_1, i_2, \ldots, i_k, \ldots, i_{l_r}). \tag{13}$$

Количество цифр в записи числа будем называть длиной числа.

Имеет место следующее утверждение:

Лемма. *Пусть заданы $l' = \max\limits_{\leqslant N}\{l_r\}$ и l — целые положительные числа. Для любого целого положительного числа не больше $i_1, i_2 \ldots i_k \ldots i_{l'}(n)$ в позиционной системе счисления с основанием n найдется равное ему число длиною не больше l в позиционной системе счисления с основанием $m = [n^{l'/l}]'$, где*

$$[n^{l'/l}]' = \begin{cases} n^{l'/l}, & \text{если } n^{l'/l} \text{ целое число} \\ [n^{l'/l}] + 1, & \text{в противном случае,} \end{cases}$$

$[n^{l'/l}]$ — целая часть числа $n^{l'/l}$.

Из справедливости леммы следует, что любое число длины не большей l' в позиционной системе счисления с основанием n может быть представлено равным ему числом длиною не больше l в позиционной системе счисления с основанием m.

Пусть R^n_m — правило перевода чисел из позиционной системы счисления с основанием n в числа позиционной системы счисления с основанием m, а R_n^m — обратное правило [3].

Тогда

$$R^n_m(i_1 i_2 \ldots i_k \ldots i_{l_r}(n)) = i'_1 i'_2 \ldots i'_k \ldots i'_l(m), \tag{14}$$

$$R^m_n(i_1' i_2' \ldots i_k' \ldots i_l'(m)) = i_1 i_2 \ldots i_k \ldots i_{l_r}(n), \tag{15}$$

где $i_1', i_2', \ldots, i_k', \ldots, i_l'$ — цифры позиционной системы счисления с основанием m и цифрами

$$0, 1, 2, \ldots, m-1 = (k, k = 0, 1, 2, \ldots, m-1). \tag{16}$$

Пусть F^m — правило, устанавливающее взаимно однозначное соответствие между символами в алфавите (4) и цифрами позиционной системы счисления с основанием m (16), аналогично правилу (11):

$$b_k \overset{F^m}{\rightleftarrows} k, \ 0 \leqslant k \leqslant m-1. \tag{17}$$

Следовательно,

$$F^m(i_1' i_2' \ldots i_k' \ldots i_l'(m)) = \sigma_l(\mathfrak{B}, i_1', i_2', \ldots, i_k', \ldots, i_l'), \tag{18}$$

$$F^m(\sigma_l(\mathfrak{B}, i_1', i_2', \ldots, i_k', \ldots, i_l') = i_1' i_2' \ldots i_k' \ldots i_l'(m). \tag{19}$$

Таким образом, вследствие (12), (14), (18) имеем

$$\sigma_{l_r}(\mathfrak{A}, i_1, i_2, \ldots, i_k, \ldots, i_{l_r}) \overset{F^n}{\rightleftarrows} i_1 i_2 \ldots i_k \ldots i_{l_r}(n) \overset{R^n_m}{=}$$

$$\overset{R^n_m}{=} i'_1 i'_2 \ldots i'_k \ldots i'_l(m) \overset{F^n}{\rightleftarrows} \sigma_l(\mathfrak{B}, i'_1, i'_2, \ldots, i'_k, \ldots, i'_l),$$

где

$$m = [n^{l'/l}]', \quad l = \max_{1 \leqslant r \leqslant N}\{l_r\}.$$

Если в качестве $K'(\mathfrak{A}, \mathfrak{B})$ использовать последовательно F^n (11), R^n_m, F^m (17), то

$$K'(\mathfrak{A}, \mathfrak{B})(\sigma_{l_r}(\mathfrak{A}, i_1, i_2, \ldots, i_k, \ldots, i_{l_r})) = \sigma_l(\mathfrak{B}, i'_1, i'_2, \ldots, i'_k, \ldots, i'_l).$$

Аналогично, вследствие (13), (15), (19), имеем

$$\sigma_l(\mathfrak{B}, i'_1, i'_2, \ldots, i'_k, \ldots, i'_l) \overset{F^m}{\rightleftarrows} i'_1 i'_2 \ldots i'_k \ldots i'_l(m) \overset{R^m_n}{=} i_1 i_2 \ldots i_k \ldots i_{l_r}(n) \overset{F^n}{\rightleftarrows}$$

$$\overset{F^n}{\rightleftarrows} \sigma_{l_r}(\mathfrak{A}, i_1, i_2, \ldots, i_k, \ldots, i_{l_r}),$$

где

$$m = [n^{l'/l}]', \quad l = \max_{1 \leqslant r \leqslant N} \{l_r\}.$$

Если в качестве $K''(\mathfrak{B}, \mathfrak{A})$ использовать последовательно F^m (17), R^m_n, F^n (11), то

$$K''(\mathfrak{B}, \mathfrak{A})(\sigma_l(\mathfrak{B}, i'_1, i'_2, \ldots, i'_k, \ldots i'_l)) = \sigma_{l_r}(\mathfrak{A}, i_1, i_2, \ldots, i_k, \ldots, i_{l_r}).$$

Тем самым теорема доказана полностью.

Назовем высказыванием в алфавите $\mathfrak{A}$ конечную совокупность слов (2) в алфавите $\mathfrak{A}$ (1), объединенную в одно слово с помощью пробелов a_0:

$$\sigma_{l_1}(\mathfrak{A}, i_1, i_2, \ldots, i_k, \ldots, i_{l_1})\, a_0 \sigma_{l_2}(\mathfrak{A}, i_1, i_2, \ldots, i_k, \ldots, i_{l_2})\, a_0 \ldots a_0 \sigma_{l_s} \times$$

$$\times (\mathfrak{A}, i_1, i_2, \ldots, i_k, \ldots, i_{l_s}).$$

Длиной высказывания назовем количество символов в нем, включая пробелы между словами,

$$l = \sum_{k=1}^{s} (l_k + 1) - 1.$$

Совокупность всевозможных высказываний из слов словаря $\Sigma_N(\mathfrak{A})$ (3) будем обозначать через $D(\Sigma_N(\mathfrak{A}))$.

Аналогично введем высказывание в алфавите $\mathfrak{B}$ (4) и $D(\Sigma_N(\mathfrak{B}))$ — множество всевозможных высказываний из слов словаря $\Sigma_N(\mathfrak{B})$ (6).

Очевидно, правила F^n (11), R^n_m, F^m (17), R^m_n можно распространить и на высказывания. Тогда доказанная теорема будет справедлива и для произвольных высказываний из $D(\Sigma_N(\mathfrak{A}))$.

Отдел по внедрению экономико-математических методов
в планирование народного хозяйства
Госплана СССР
Москва

Поступило
5 III 1970

ЦИТИРОВАННАЯ ЛИТЕРАТУРА

[1] А. И. Мальцев, Алгоритмы и рекурсивные функции, «Наука», 1965. [2] Л. Н. Королев, ДАН, 113, № 4 (1957). [3] А. И. Китов, Н. А. Криницкий, Электронные цифровые машины и программирование, М., 1961.