K. I. Kurbakov, An addressing method using condensed word code as memory addresses, *Dokl. Akad. Nauk SSSR*, 1965, Volume 163, Number 4, 841–844

Доклады Академии наук СССР
1965. Том 163, № 4

КИБЕРНЕТИКА И ТЕОРИЯ РЕГУЛИРОВАНИЯ

К. И. КУРБАКОВ

СПОСОБ АДРЕСАЦИИ, ИСПОЛЬЗУЮЩИЙ СЖАТЫЕ КОДЫ СЛОВ В КАЧЕСТВЕ АДРЕСОВ ПАМЯТИ

(Представлено академиком В. М. Глушковым 19 I 1965)

В работах ([1–3]) исследуется возможность непосредственного преобразования исходной информации в адрес запоминающего устройства (ЗУ) с произвольным обращением к его любой ячейке. Данная работа относится также к этому направлению.

В ([4]) описан способ сжатия кодов слов исходного словаря (текста), который заключается в том, что код каждой последующей буквы слова * сдвигается относительно кода предыдущей буквы в сторону старших разрядов на один разряд и суммируется по модулю 2 в каждом разряде, при этом коды букв выбираются с учетом вероятностно-статистических характеристик в языке. В результате такого преобразования получается сжатый код слова из n разрядов.

С целью уменьшения n для длинных слов, имеющих длину сжатого кода слова больше некоторой заданной величины, можно коды конечных букв слова складывать по модулю 2 в каждом разряде без сдвига относительно некоторой предыдущей буквы, т. е. сдвиг осуществлять k раз:

$$k_{\text{ф}} = n - m, \tag{1}$$

где $k_{\text{ф}}$ — фиксированное число сдвигов.

Например, пусть $n = 11$ двоичным разрядам и коды букв m соответствуют тем кодовым комбинациям, которые приведены ниже в примере, тогда сжатие слова ГАЗЕТА произойдет следующим образом:

```
                1 1 1 1 0 0 0 0      А
                1 1 0 1 1 0 0 0      Т
          +     0 0 0 1 0 1 1 1      Е
(по мод. 2)       0 1 1 0 0 0 1 1    З
                    1 1 1 1 0 0 0 0  А
                      0 1 0 0 0 1 1 1Г
              ---------------------
              0 0 1 1 1 0 1 0 0 1 1
```

L, $n \leftarrow \uparrow$

Таким образом, при $n = 11$ двоичным разрядам, $m = 8$ двоичным разрядам для слова ГАЗЕТА длиной $L = 6$ букв необходимо сделать всего три сдвига относительно первой буквы этого слова; остальные буквы (Т и А) складываются по мод. 2 в каждом разряде без сдвига относительно буквы, сдвинутой последней (Е).

Для разделения неоднозначности сжатия, возникающей в процессе пословного сжатия информации по описанному способу, могут быть применены различительные признаки, которые определяются непосредственно из исходного слова. Код признаков слова вырабатывается в блоке сжатия одновременно с сжатием кода слова.

---

* Под словом исходного словаря понимается некоторый набор букв, ограниченный пробелом, а под буквой — любой символ алфавита в марковском смысле.

В процессе пословного кодирования описанным способом осуществляется пословное сжатие информации, и в результате этого преобразования вырабатывается случайное число, которое может быть использовано в качестве адреса исходного слова. Основная цель кодирования символов алфавита с учетом вероятностно-статистических характеристик словаря (текста) и применения этого вида сжатия кодов слов заключается в том, что необходимо получить наиболее равномерное распределение сжатых кодов слов в некотором заданном интервале

$$M = 0 - 2^n, \tag{2}$$

где M — количество адресов в конкретном ЗУ, n — количество двоичных разрядов в сжатом коде слова, и использовать сжатые коды слов в качестве адресов к памяти с произвольным обращением.

Суть способа адресации, использующего сжатые коды слов в качестве адресов памяти, заключается в следующем ([5]). Код преобразованного слова, получаемый на выходе блока сжатия или вырабатываемый программным способом, состоит из двух частей: а) n_i — сжатого кода слова, т. е. адреса исходного слова; б) t_i — кода признаков исходного слова.

Полный код преобразованного слова $(n_i + t_i)$ однозначно характеризует любое слово во всем используемом множестве N исходных слов (словаре), при этом

$$2^n \geqslant N, \quad 2^{n+t} \gg N. \tag{3}$$

Множество преобразованных слов N' представлено двумя видами слов: а) слова с неповторяющимся для данного словаря адресом n_i и б) слова с k-кратным повторением для данного словаря адресом n_i, т. е. имеются группы неоднозначно сжатых кодов слов.

Слова любой группы неоднозначности сжатия η_i ($i = 1, 2, \ldots, k$, где k — последнее слово в группе) разделяются между собой признаками слова t. Все слова внутри каждой группы неоднозначности сжатия связываются обычной переадресацией. В связи с этим, по каждому адресу n_i хранится код переадресации $\gamma_i (\equiv n_{i+j})$ к следующему слову этой же группы неоднозначности сжатия.

Рассмотрим кратко алгоритм поиска в автоматическом словаре, который достаточно просто реализуется программным способом на существующих ЭВМ. После создания преобразованного кода исходному слову $(n_i + t_i)$ обращение к ЗУ с произвольным доступом к каждой ячейке памяти осуществляется по коду адреса n_i, т. е. по части преобразованного кода исходного слова. В результате по этому адресу считывается «число», которое состоит из трех частей α, β и γ, где $\alpha_i (\equiv n_i)$ — информация (И), связанная с исходным словом (например, эквивалент перевода слова с одного языка на другой с сопровождающей его грамматической и тому подобной информацией); $\beta_i (\equiv t)$ — код признаков исходного слова; $\gamma_i (\equiv n_{i+j})$ — код переадресации или код, используемый для изменения адреса n_i.

До выдачи информации по запрошенному адресу осуществляется сравнение признаков исходного слова с признаками слова, хранимого в ЗУ и найденного по адресу $\alpha_i (\equiv n_i)$. Если признаки совпадают ($t_{\text{псх}.\,i} = t_{\text{ЗУ},\,i}$), то выдается значение И_{α_i}. Если же $t_{\text{исх},\,i} \neq t_{\text{ЗУ},\,i}$, то выдача значения И_{α_i} блокируется и осуществляется переадресация по адресу, указанному в $\gamma_i (\equiv n_{i+j})$. Процесс сравнения признаков слова внутри η_i-й группы неоднозначности сжатия кодов слов продолжается до тех пор, пока не будет найден эквивалент исходному слову * или не получен

---

* Под $\alpha_i (\equiv n_i)$ при $t_{\text{исх}} = t_{\text{ЗУ}}$ подразумевается также адресная отсылка к некоторой части оперативного ЗУ или к внешним ЗУ (например, накопителям типа магнитных лент, дисков и барабанов).

сигнал «Такого слова в словаре нет». Признаком того, что исходного слова в словаре нет, является отсутствие кода (нулевая группа или другой специальный признак) в разрядах переадресации $\gamma_{i=k}(\equiv 0)$ последнего (k-го) слова группы неоднозначности сжатия (группы слов с переадресацией).

Среднее время поиска одного слова ($T_{\text{ср. п}}$) в машинном словаре по сжатому коду исходного слова для рассматриваемого способа адресации определяется средней длиной группы неоднозначности сжатия η_i и в общем виде определяется как отношение суммарного количества циклов *, затраченных на поиск всех слов словаря (без переадресации, с одной, двумя и т. д. переадресациями), к количеству всех слов в словаре, т. е.

$$T_{\text{ср.п.}} = \left[\sum_{1}^{\eta_k} \frac{\eta_i(\eta_i+1)}{2\eta_i!}\right] \Big| \left[\sum_{1}^{\eta_k} \frac{\eta_i}{\eta_i!}\right]. \quad (4)$$

Поскольку отношение $\frac{1}{\eta_i!}$ при $\eta_k = 10$ практически мало́, то, ограничиваясь значением $\eta_{i=1,\,2,\,\ldots,\,k} = 1, 2, \ldots, 10$, получаем

$$T_{\text{ср. п}} = 1{,}500 \text{ цикла.}$$

Использование сжатых кодов слов в качестве адресов памяти позволяет значительно уменьшить среднее время поиска одного слова в словаре по сравнению с традиционными способами поиска. Это достигается благодаря тому, что поиск в словаре по рассматриваемому способу сводится в основном к обращению в словарь по однозначно сжатым словам (адресам), а поиск внутри неоднозначности сжатия слов (т. е. с переадресацией) в среднем для всего словаря и рассматриваемого способа преобразования слов в адреса памяти незначителен.

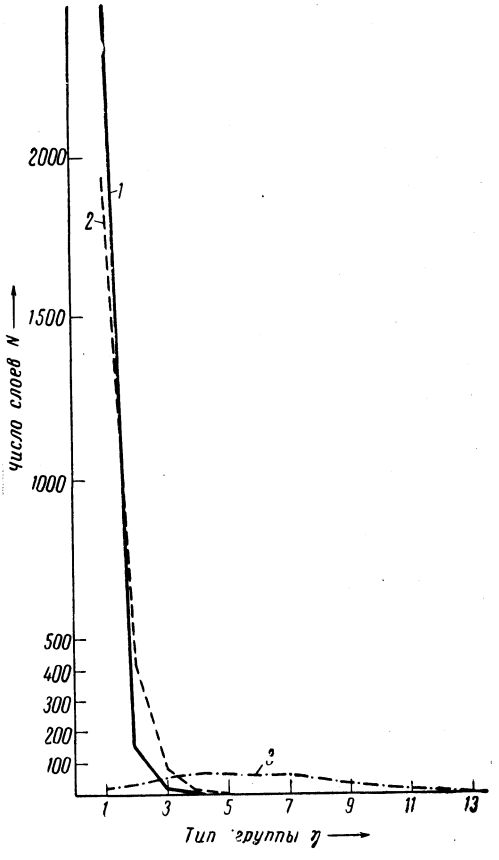

Рис. 1. Распределение неоднозначности сжатия кодов слов. *1* — $n = n_{\max} = 30$ двоичных разрядов; *2* — $n_{\text{опт}} = 12$ двоичных разрядов; *3* — $n = n_{\min} = 8$ двоичных разрядов

На рис. 1 показано распределение неоднозначности сжатия при почти оптимальном варианте кодирования букв, длине кода буквы $m = 8$ двоичным разрядам и объеме словаря в $N = 3000$ словоформ для трех фиксированных точек сжатия кодов слов исходного словаря:

$$1)\ 2^{n\max} \gg N;\ 2)\ 2^{n\text{опт}} \geqslant N;\ 3)\ 2^{n\min} \ll N. \quad (5)$$

Как видно из рисунка, для случая 2) 80,4% всех слов словаря N однозначно преобразуется в адрес к ЗУ и лишь 19,6% падает на неоднозначность сжатия. Для случая 2) $T_{\text{ср. п}} = 1{,}233$ цикла.

---

* Под циклом обращения к словарю или поиска в общем случае понимается совокупность арифметических и логических операций, необходимых для одной идентификации (одного сравнения заданного кода слова с кодом некоторого слова в машинном словаре).

Общее время поиска одного слова в машинном словаре по данному способу также мало по сравнению с $T_{общ}$ в традиционных способах поиска по словарю (например, при способе поиска делением словаря пополам, при использовании способа разделителей и т. п.) и равно

$$T_{общ} = T_{сж} + T_{ср.\,п}, \tag{6}$$

где $T_{сж}$ — время сжатия кода слова (время преобразования кода исходного слова в адрес), которое равно в среднем 8—9 операциям сдвига и 8—9 операциям сложения по модулю 2, если данное преобразование осуществляется программным способом и если считать среднюю длину словоформы равной $L_{ср} = 8—9$ буквам.

Поступило
24 XII 1964

ЦИТИРОВАННАЯ ЛИТЕРАТУРА

[1] W. W. Peterson, IBM J. Res. and Developm., 1, № 2, 130 (1957). [2] L. R. Johnson, Comm. ACM, 4, № 5, 218 (1961). [3] G. Schay, N. Raver, IBM J. Res. and Developm., 7, № 2, 121 (1963). [4] Р. В. Смирнов, К. И. Курбаков, Авт. свид. № 149264, 1961; Бюлл. изобретений, № 15 (1962). [5] К. И. Курбаков, Р. В. Смирнов, Авт. свид. № 153800, 1962; Бюлл. изобретений, № 7 (1963).