Н.П. БРУСЕНЦОВ

О ВЫЧИТАНИИ И ОКРУГЛЕНИИ ЧИСЕЛ В ПОЗИЦИОННЫХ СИСТЕМАХ СЧИСЛЕНИЯ С ПОЛОЖИТЕЛЬНЫМ ОСНОВАНИЕМ

Рассмотрим p-ичную позиционную систему счисления с цифрами a_1, a_2, a_p, которым в порядке их старшинства соответствуют последовательные целочисленные значения, причем наибольшее из этих значений, соответствующее старшей цифре $S_p = S\ (a_p)$ меньше целого числа p, являющегося основанием системы счисления.

В рассматриваемой системе слово вида

$$x_m\ x_{m-1}\ \dots\ x_2\ x_1\ x_0,\ x_{-1}\ x_{-2}\ \dots\ x_1\ \dots\ x_k \tag{1}$$

в котором символами x_1 являются цифры a_1, a_2, ... a_p, интерпретируется как число x, выражающееся суммой

$$x = \sum_{i=-k}^{m} S(x_i)p^i. \tag{2}$$

Слово (1) называется прямым кодом числа x.

Обратным кодом числа будем называть код, полученный заменой всех цифр прямого кода этого числа по формуле

$$[a_i]_\text{обр} = a_{p+1-i}, \tag{3}$$

где a_i – цифра в прямом коде числа; $[\ ]_\text{обр}$ – обозначает операцию обращения; a_{p+1-i} – цифра обратного кода, соответствующая цифре a_i.

Цифры, связанные формулой (3), являются взаимообратными, т. е. если $[a_i]_\text{обр} = a_j$, то $[a_j]_\text{обр} = a_i$.

Действительно, если $[a_i]_\text{обр} = a_j$, то, согласно (3), $a_j = a_{p+1-i}$ и, снова, согласно (3), $[a_j]_\text{обр} = [a_{p+1-i}]_\text{обр} = a_i$. Это означает, в частности, что в результате обращения, произведенного четное число раз, получается исходная цифра.

Обращение цифры равносильно вычитанию значения этой цифры из некоторого фиксированного для данной системы счисления целого числа q. Действительно, в системе счисления с основанием p и значением старшей цифры S_p значение цифры a_i будет равно

$$S_i = S_p - p + i, \tag{4}$$

а значение цифры, обратной a_i, $[a_i]_\text{обр} = a_{p+1-i}$, будет равно

$$S_{p+1-i} = S_p + 1 - i. \tag{5}$$

Сумма этих значений

$$S_i + S_{p+|1-\text{i}} = 2S_p - p + 1 \tag{6}$$

не зависит от номера цифры i, т. е. является фиксированным числом

$$q = 2S_p - p|+|1.$$

Число q представляет собой удвоенное среднее арифметическое значений S_i, приписанных цифрам системы счисления,

$$\text{q} = \frac{2}{p} \sum_{i=k-2}^{0} S_i. \tag{7}$$

В системе с основанием p выбор q, $|q| < p$ определяет значения всех цифр. В частности, значение старшей цифры

$$S_p = \frac{p+q-1}{2} \tag{8}$$

и значение, максимальное по абсолютной величине.

$$|S|_{\max} = \frac{p+|q|-1}{2}. \tag{9}$$

В системах с неотрицательными значениями цифр число q совпадает со значением старшей цифры $q = S_p$; в системах с симметричным относительно нуля расположением значений цифр $q = 0$.

Обращение кодов позволяет, как известно, существенно упростить реализацию вычитания, представив отрицательные числа дополнениями до некоторого фиксированного числа a

$$x-y = [\ x+(a-y)\ ]-a. \tag{10}$$

Практическая ценность этой формулы зависит от того, насколько простыми будут операции $a-y$ и $[\ ]-a$, т. е. от того, найдется ли такое значение числа a, при котором эти операции будут простыми.

Операция $a-y$ сводится к обращению кода y, если число а таково, что во всех его разрядах содержится цифра, обладающая значением q. Операция $[\ ]-a$ будет простой, если a=0 или $a=p^{m+r}$, где m – номер старшего разряда в представлении чисел, с которыми производятся вычисления; r – положительное целое число; $r \geq 1$.

Поскольку число, в котором все цифры одинаковы, не может быть целой степенью основания системы счисления, то оба предъявляемые к числу a требования выполняются совместно в единственном случае: при q=0, т. е. при симметричном относительно нуля расположении значений цифр. В этом случае вычитание сводиться к сложению с предварительным обращением кода вычитаемого.

В системах счисления с неотрицательными значениями цифр, в том числе в двоичной системе с цифрами 0, 1 и в десятичной системе с цифрами 0, 1, ... , 9 благодаря тому, что $q= S_p$ число, образованное повторением цифры со значением q, лишь на единицу младшего разряда отличается от целой степени основания системы. В связи с этим в качестве числа a с равным успехом используется или целая степень основания системы (дополнительный код), или число, образованное повторением старшей цифры во всех разрядах (обратный код). В обоих случаях реализация вычитания связана с необходимостью производить дополнительное сложение чисел с единицей младшего разряда.

В других системах счисления ($q \neq 0$, $q \neq S_p$) рассмотренный способ упрощения операции вычитания применить не удастся.

Округлить до n разрядов число x, представленное более чем n разрядами, значит заменить это число таким числом x', у к оторого в младших разрядах, расположенных правее n старших разрядов числа x, содержатся нули, а цифры остальных разрядов выбраны так, чтобы абсолютная величина разности $x-x'$ была минимальной.

Рассмотрим округление действительного числа x, точное значение которого выражается в p-ичной системе счисления суммой

$$x = \sum_{i=-\infty}^{m} S(x_i)\cdot p^i, \tag{11}$$

где m − номер самого старшего разряда числа, отсчитанный относительно нулевого ($i = 0$) разряда.

Число x', представленное n старшими разрядами числа x с сохранением находящихся в этих разрядах цифр

$$x' = \sum_{i=m-n+1}^{m} S(x_i)\cdot p^i, \tag{12}$$

отличается от числа x на величину

$$x-x' = \sum_{i=-\infty}^{m-n} S(x_i)\cdot p^i. \tag{13}$$

Эта величина будет наибольшей по абсолютному значению в том случае, когда во всех отброшенных разрядах содержится цифра, обладающая максимальным абсолютным значением $|\,S\,|_{\max}$,

$$|\,x-x'\,|_{\max} = |\,S\,|_{\max} \sum_{i=-\infty}^{m-n} p^i = |\,S\,|_{\max} \frac{p^{m+n+1}}{p-1}. \tag{14}$$

По отношению к единице младшего из сохраняемых в (12) разрядов это составит

$$\frac{|\,x-x'|_{\max}}{p^{m-n-1}} = \frac{|\,S\,|_{\max}}{p-1}. \tag{15}$$

Таким образом, число (12) непременно будет правильным округлением числа (11), если

$$\frac{|\,S\,|_{\max}}{p-1} \le 0{,}5. \tag{16}$$

В противном случае может возникнуть необходимость корректировки, приводящей, вообще говоря, к изменению цифр во всех разрядах числа (12), включая (m+1)-й разряд, расположенный левее m-го разряда.

Подстановка в (16) выражения (9) для $|\,S\,|_{\max}$ дает формулу

$$\frac{p+|\,q\,|-1}{p-1} \le 1, \tag{17}$$

которая позволяет заключить о возможности правильного округления простым отбрасыванием младших разрядов в различных системах счисления. Так как $p > 1$, то неравенство (17) удовлетворяется в единственном случае при $q = 0$. Таким образом, в системах счисления с положительным основанием только в случае симметричного относительно нуля расположения значений цифр округление сводится к простому отбрасыванию младших разрядов числа.

Статья поступила в редакцию 29 февраля 1968 г.