Математика и её приложения: ЖИМО. 2008. Выпуск 1 (5). С. 43 – 52.

УДК 512.54

А. А. Толстопятов[1]

Возможность кодирования поля кратности одним числом

Ключевые слова: поле кратности, стандартная форма, таблица перестановок.

Рассматривается возможность кодирования поля кратности не двумя числами — номером в таблице стандартных форм и номером в таблице перестановок отдельной стандартной формы, а одним — номером в таблице, объединяющей все стандартные формы и их перестановки. Сравниваются длины кодов поля кратности при этих двух способах кодирования. Рассматриваются асимптотики длин кодов этих двух способов.

We consider the possibility of coding the multiplicity field by one integer — the number in the table that unites all standard forms and their swaps. The other method of coding the multiplicity field consists of using two integers — the number in the table of standard forms and the number in the table of swaps some standard form. We compare code lengths of the multiplicity field for this methods; also we consider asimptotics for code lengths.

Предложенный в [3] подход к сжатию информации требует построения кода отдельных буферов, на которые разбит файл. Этот код состоит из трех полей — принадлежности, кратности и порядка. Предложенные в [5, 11] алгоритмы построения кода поля кратности дают код, состоящий из двух чисел — номера в таблице стандартных форм и номера в таблице перестановок данной стандартной формы. Вычисление последнего номера требует обращения к алгоритму вычисления номера в таблице порядка [12, 2]. Это делает алгоритм построения кода поля порядка достаточно сложным. Поэтому появляется идея объединить таблицы стандартных форм и перестановок в одну таблицу. Тогда код поля кратности будет содержать только одно число. Делать это имеет смысл только в случае, если длина так построенного кода не будет превосходить длину кода из двух чисел. Этот вопрос и обсуждается в настоящей статье.

1. Постановка задачи

Буфер состоит из кортежей длины n. Код поля принадлежности восстанавливает эти кортежи, но без учета повторов. Поэтому, если в буфер первый кортеж входит n_1 раз, второй — n_2, ..., s-й — n_s раз, то числа повторов n_1, n_2, ..., n_s — это та информация, которая должна восстанавливаться по коду поля порядка. Пусть m — число кортежей в буфере,

[1] Ивановский государственный университет; E-mail: khash2@mail.ru. Работа выполнена при финансовой поддержке РФФИ (проект 07-07-00155).

а s — число разных кортежей. Тогда, если числа повторов упорядочить лексикографически:

$$n_1 \le n_2 \le \cdots \le n_s \tag{1}$$

и при заданных m и s составить таблицу всех разложений числа m в s слагаемых, упорядоченных согласно (1), то мы получим таблицу стандартных форм. По заданным числам повторов номер в этой таблице вычисляется однозначно, так же как и по номеру однозначно восстанавливаются числа повторов [5, 11]. Обозначим этот номер через N_1. Как показано в [4], длина таблицы стандартных форм есть D^s_{m-s}. Эта величина может быть вычислена, например, с использованием рекуррентных формул:

$$\begin{array}{lll} 1.\ D^s_n & = D^s_{n-s} + D^{s-1}_n, & s \le n, \\ 2.\ D^s_n & = D^n_n, & s > n, \\ 3.\ D^s_0 & = 1, & \\ 4.\ D^1_n & = 1, & \\ 5.\ D^2_n & = \left[\frac{n}{2}\right] + 1. & \end{array} \tag{2}$$

Поэтому

$$N_1 \le D^s_{m-s}. \tag{3}$$

Однако, в буфере числа повторов $n_1, n_2, \ldots, n_s$ могут не удовлетворять условию (1), а быть расположенными в произвольном порядке. Поэтому для каждой стандартной формы необходимо построить таблицу перестановок. Номер перестановки в этой таблице обозначим через N_2. Если среди чисел повторов $n_1, n_2, \ldots, n_s$ есть l_1 одинаковых, l_2 одинаковых, но отличных от первых, $\ldots$, l_p одинаковых, но отличных от первых, вторых, $\ldots$, $(p-1)$-х, причем

$$\sum_{i=1}^{p} l_i = s, \tag{4}$$

то таблица перестановок будет иметь длину σ, равную

$$\sigma = \frac{s!}{\prod\limits_{i=1}^{p} l_i!}. \tag{5}$$

Тогда справедливо, что

$$N_2 \le \sigma. \tag{6}$$

Таким образом, код поля кратности имеет следующий шаблон

N_1	N_2	(7)

и содержит два числа.

Теперь откажемся от лексикографического упорядочения (1), т. е. от таблицы стандартных форм, и составим таблицу всех разложений числа m в s слагаемых, упорядоченную в соответствии со следующими правилами.

1. Первое разложение содержит $s-1$ первых слагаемых, равных 1, а s-е слагаемое равно $m-s+1$.

2. Последовательно увеличиваем $(s-1)$-е слагаемое на единицу, одновременно уменьшая s-е слагаемое на единицу, пока $(s-1)$-е слагаемое не станет равным $m-s+1$, а s-е — единице.

3. Увеличиваем $(s-2)$-е слагаемое до двойки, $(s-1)$-е слагаемое принимаем равным единице, а s-е слагаемое — $m-s$.

4. Последовательно увеличиваем $(s-1)$-е слагаемое на единицу, одновременно уменьшая s-е слагаемое на единицу, пока $(s-1)$-е слагаемое не станет равным $m-s$, а s-е — единице.

5. Увеличиваем $(s-2)$-е слагаемое до тройки, $(s-1)$-е слагаемое принимаем равным единице, а s-е слагаемое — $m-s-1$.

6. Продолжаем эту процедуру до тех пор, пока 1-е слагаемое не станет равным $m-s+1$, а 2-е, 3-е, $\ldots$, s-е равными единице.

В результате получится таблица чисел порядка, включающая в себя не только все возможные наборы чисел порядка n_k, $k=1, 2, \ldots, s$, удовлетворяющие условию

$$\sum_{k=1}^{s} n_k = m, \tag{8}$$

но и расположенные во всех возможных порядках. Например, строки в этой таблице, которые обозначим через N, будут однозначно вычисляться по числам порядка $n_1, n_2, \ldots, n_s$, без условия (1), а по N однозначно вычисляются сами числа порядка. Тогда код поля кратности будет содержать только один номер N.

Длину таблицы, построенной по описанному выше алгоритму, нетрудно вычислить. Действительно, число различных разложений (8) равно C_{m-1}^{s-1} т. е. количеству наборов s чисел n_k из совокупности $\{1, 2, \ldots, n-s+1\}$ при ограничении (8). Значит рассматриваемая таблица содержит C_{m-1}^{s-1} строк, и число N удовлетворяет неравенству

$$N \le C_{m-1}^{s-1}, \tag{9}$$

а шаблон поля кратности будет

$$\boxed{N}. \tag{10}$$

Задача заключается в том, чтобы ответить на вопрос, длина какого шаблона (7) или (10) будет меньше при заданных m и s.

2. Сравнение длин кодов поля кратности

При кодировании поля кратности двумя числами N_1 и N_2 длины кода будут складываться из длин этих чисел L_1 и L_2. Длины L_1 и L_2 можно оценить по максимальным значениям чисел N_1 и N_2, и так как[2]

$$L_1 = \{\log_2(N_1)\}, \tag{11}$$

то вследствие (3) будем иметь:

$$L_1 = \{\log_2(D^s_{m-s})\}. \tag{12}$$

Формула (5) справедлива для отдельной стандартной формы, причем для разных стандартных форм σ будет разной. Поэтому, вводя обозначение $\sigma_{\max}$ согласно формуле

$$\sigma_{\max} = \max_{k=1}^{D^s_{m-s}} \sigma_k, \tag{13}$$

получим с учетом (6), что

$$L_2 = \{\log_2 \sigma_{\max}\}. \tag{14}$$

При заданных m и s число σ из (5) будет тем больше, чем меньше знаменатель правой части (5); а знаменатель тем меньше, чем больше разных n_k в разложении (8). Так как $m \geq s$, то нетрудно получить минимальное значение m, при котором возможно все l_i сделать равными единице, а значит и все знаменатели в (5) тоже будут равны единице. Действительно, если m удовлетворяет неравенству

$$m \geq \frac{s(s+1)}{2}, \tag{15}$$

то m можно записать в виде

$$m = \frac{s(s+1)}{2} + k, \tag{16}$$

где $k > 0$. Тогда разложение (8) вида

$$m = 1 + 2 + 3 + \cdots + (s-1) + (s+k) \tag{17}$$

будет содержать s слагаемых $n_1 = 1,\ n_2 = 2,\ \ldots,\ n_{s-1} = s-1,\ n_s = s+k$, удовлетворяющих строгим неравенствам

$$n_1 < n_2 < \cdots < n_{s-1} < n_s. \tag{18}$$

[2] $\{x\}$ — наименьшее целое число, превосходящее x.

Вследствие (18) все числа повторов будут разные, а значит

$$l_1 = l_2 = \cdots = l_s = 1,$$

т. е. при выполнении (15)

$$\sigma_{\max} = s!. \tag{19}$$

Нетрудно показать, как будут меняться $\sigma_{\max}$, если (15) не выполняется. Например, если имеет место неравенство

$$\frac{s(s-1)}{2} < m < \frac{s(s+1)}{2}, \tag{20}$$

то

$$\sigma_{\max} = \frac{s!}{2}. \tag{21}$$

Однако, поскольку полная классификация соотношений между m и s занимает очень много места[3], то достаточно ограничиться оценкой (19), и тогда из (14) и (19) будем иметь:

$$L_2 = \{\log_2 s!\}. \tag{22}$$

Обозначая длину кода поля принадлежности при кодировании первым способом (т. е. двумя числами N_1 и N_2) через $\mathcal{L}_1 = L_1 + L_2$, из (12) и (22) получим, что

$$\mathcal{L}_1 = \{\log_2 D^s_{m-s}\} + \{\log_2 s!\}. \tag{23}$$

Для второго способа кодирования (т. е. одним числом N_1) обозначим длину кода через $\mathcal{L}_2$; из (9) найдем, что

$$\mathcal{L}_2 = \{\log_2 C^{s-1}_{m-1}\}. \tag{24}$$

Вопрос заключается в том, какое из следующих неравенств и при каких m и s будет выполняться:

$$\{\log_2 D^s_{m-s}\} + \{\log_2 s!\} > \{\log_2 C^{s-1}_{m-1}\} \tag{25}$$

или

$$\{\log_2 D^s_{m-s}\} + \{\log_2 s!\} < \{\log_2 C^{s-1}_{m-1}\}. \tag{26}$$

[3]Требует рассмотрения многих случаев, но при этом мало влияет на величину $\sigma_{\max}$, т. к. во всех этих случаях $\sigma_{\max} < s!$, а в оценке L_2 согласно (14) $\sigma_{\max}$ стоит под знаком логарифма.

3. Асимптотика длин $\mathcal{L}_1(m,s)$ и $\mathcal{L}_2(m,s)$

Неравенства (25) или (26) будут выполняться для определенных значений параметров m и s. Найти интервалы изменения m и s, внутри которых выполняется (25) или (26), можно или последовательно придавая m и s определенные значения, для которых левая и правая части могут быть вычислены, и сравнивая эти части,[4] или рассматривая ассимптотические случаи, когда $s \sim m$ (в буфере мало повторов) и $s << m$ (в буфере много повторов).

Асимптотика величины $\log_2 D^s_{m-s}$ была получена методом перевала в [4] и имеет следующий вид:

$$\begin{array}{ll} \log_2 D^s_{m-s} \approx 2,82 \cdot (sm)^{1/2} & \text{при } m >> s >> 1\ , \\ \log_2 D^s_{m-s} \approx 2,56 \cdot (m)^{1/2} & \text{при } m \sim s >> 1\ . \end{array} \tag{27}$$

Чтобы найти асимптотическое поведение $\log_2 s!$ и $\log_2 C^{s-1}_{m-1}$ воспользуемся формулой Стирлинга, справедливой для $n >> 1$:

$$n! \approx (2\pi n)^{1/2} \left(\frac{n}{e}\right)^n . \tag{28}$$

В результате найдем, что

$$\begin{array}{ll} \log_2 s! \approx s \cdot \log_2 s & \text{при } m >> s >> 1\ , \\ \log_2 s! \approx m \cdot \log_2 m & \text{при } m \sim s >> 1\ ; \end{array} \tag{29}$$

$$\begin{array}{ll} \log_2 C^{s-1}_{m-1} \approx s \cdot \log_2 m & \text{при } m >> s >> 1\ , \\ \log_2 C^{s-1}_{m-1} \approx (m-s) \cdot \log_2 m & \text{при } m \sim s >> 1\ . \end{array} \tag{30}$$

Из (27), (29), (30) найдем, что

$$\begin{array}{ll} \mathcal{L}_1 \approx 2,82 \cdot (sm)^{1/2} + s \log_2 s & \text{при } m >> s >> 1\ , \\ \mathcal{L}_1 \approx 2,56 \cdot (m)^{1/2} + m \log_2 m & \text{при } m \sim s >> 1\ ; \end{array} \tag{31}$$

и

$$\begin{array}{ll} \mathcal{L}_2 \approx s \log_2 m & \text{при } m >> s >> 1\ , \\ \mathcal{L}_2 \approx (m-s) \log_2 m & \text{при } m \sim s >> 1\ . \end{array} \tag{32}$$

В случае, когда $m >> s >> 1$, рассмотрим разность $\mathcal{L}_1 - \mathcal{L}_2$, которая согласно (31) и (32) равна:

$$\mathcal{L}_1 - \mathcal{L}_2 \approx 2,82 \cdot (sm)^{1/2} - s \log_2 \frac{m}{s} . \tag{33}$$

Поскольку $m >> s >> 1$ и $m^{1/2}$ растет быстрее $\log_2 m$, то ведущий член в (33) есть $(sm)^{1/2}$ и поэтому

$$\mathcal{L}_1 - \mathcal{L}_2 \approx 2,82 \cdot (sm)^{1/2} > 0 . \tag{34}$$

[4] На малоразмерном примере это будет рассмотрено в следующем разделе работы.

Из (34) следует, что

$$\mathcal{L}_1 > \mathcal{L}_2 \ . \tag{35}$$

Таким образом, *если* $m >> s >> 1$, *то кодирование одним числом оказывается более экономичным.*

В случае $m \sim s >> 1$ из (31) и (32) получим, что

$$\mathcal{L}_1 - \mathcal{L}_2 \approx 2,56 \cdot (m)^{1/2} + s \log_2 m > 0 \ , \tag{36}$$

т. е. и в этом случае справедливо неравенство (35).

Таким образом, рассмотрение ассимптотического поведения длин кодов поля кратности показывает, что *второй способ кодирования является более экономичным.* Поэтому есть смысл построить соответствующие алгоритмы для вычисления номера N, отказавшись от условия (1). Чтобы убедиться в этом выводе, рассмотрим малоразмерный пример.

4. Пример

Рассмотрим пример с частными значениями параметров m и s, чтобы соответствующие таблицы можно было построить не как виртуальные, а как реальные. Возьмем $m = 10,\ s = 4$. Тогда длина таблицы стандартных форм будет равна $D^s_{m-s} = D^4_6$. По рекуррентной формуле (2) найдем, что $D^4_6 = D^4_2 + D^3_6 = D^2_2 + D^3_0 + D^3_3 + D^2_6 = 9$.

Таблица 1. Длина кода поля кратности при кодировании двумя числами при $m = 10,\ s = 4$.

№	Стандартная форма	$\{\log_2 N_1\}$	σ	$\{\log_2 \sigma\}$	$\mathcal{L}_1$
1	10=1+1+1+7	1	4	2	3
2	10=1+1+2+6	1	12	4	5
3	10=1+1+3+5	2	12	4	6
4	10=1+1+4+4	2	6	3	5
5	10=1+2+2+5	3	12	4	7
6	10=1+2+3+4	3	24	5	8
7	10=1+3+3+3	3	4	2	5
8	10=2+2+2+4	3	4	2	5
9	10=2+2+3+3	4	6	3	7

Отметим, что при вычислении $\mathcal{L}_1$ по формуле (23) получилось бы следующее:

$$\mathcal{L}_1(10,4) = \{\log_2 D^3_6\} + \{\log_2 4!\} = 4 + 5 = 9, \tag{37}$$

т. е. приведенная выше оценка для данного примера оказывается чуть выше максимума. В (37) введено обозначение

$$\mathcal{L}_1(10,4) = \mathcal{L}_1(m,s) \text{ при } m = 10,\ s = 4. \tag{38}$$

В случае кодирования одним числом длина кода будет

$$\mathcal{L}_2(10,4) = \{\log_2 C_9^3\} = \{\log_2 84\} = 7, \tag{39}$$

что меньше, чем $\mathcal{L}_1(10,4)$ из (37), и даже меньше, чем максимальное значение из Табл. 1, равное 8.

Чтобы проиллюстрировать на разбираемом примере однозначность построения таблицы без условия (1), алгоритм которого был рассмотрен в первом разделе статьи, построим эту таблицу для $m = 10,\ s = 4$.

Таблица 2. Разложения (8) при кодировании поля порядка одним числом для $m = 10,\ s = 4$

1. 10=1+1+1+7	7. 10=1+1+7+1	13. 10=1+2+6+1
2. 10=1+1+2+6	8. 10=1+2+1+6	14. 10=1+3+1+5
3. 10=1+1+3+5	9. 10=1+2+2+5	15. 10=1+3+2+4
4. 10=1+1+4+4	10. 10=1+2+3+4	16. 10=1+3+3+3
5. 10=1+1+5+3	11. 10=1+2+4+3	17. 10=1+3+4+2
6. 10=1+1+6+2	12. 10=1+2+5+2	18. 10=1+3+5+1
19. 10=1+4+1+4	41. 10=2+3+2+3	63. 10=3+4+2+1
20. 10=1+4+2+3	42. 10=2+3+3+2	64. 10=3+5+1+1
21. 10=1+4+3+2	43. 10=2+3+4+1	65. 10=4+1+1+4
22. 10=1+4+4+1	44. 10=2+4+1+3	66. 10=4+1+2+3
23. 10=1+5+1+3	45. 10=2+4+2+2	67. 10=4+1+3+2
24. 10=1+5+2+2	46. 10=2+4+3+1	68. 10=4+1+4+1
25. 10=1+5+3+1	47. 10=2+5+1+2	69. 10=4+2+1+3
26. 10=1+6+1+2	48. 10=2+5+2+1	70. 10=4+2+2+2
27. 10=1+6+2+1	49. 10=2+6+1+1	71. 10=4+2+3+1
28. 10=1+7+1+1	50. 10=3+1+1+5	72. 10=4+3+1+2
29. 10=2+1+1+6	51. 10=3+1+2+4	73. 10=4+3+2+1
30. 10=2+1+2+5	52. 10=3+1+3+3	74. 10=4+4+1+1
31. 10=2+1+3+4	53. 10=3+1+4+2	75. 10=5+1+1+3
32. 10=2+1+4+3	54. 10=3+1+5+1	76. 10=5+1+2+2
33. 10=2+1+5+2	55. 10=3+2+1+4	77. 10=5+1+3+1
34. 10=2+1+6+1	56. 10=3+2+2+3	78. 10=5+1+1+2
35. 10=2+2+1+5	57. 10=3+2+3+2	79. 10=5+2+2+1
36. 10=2+2+2+4	58. 10=3+2+4+1	80. 10=5+3+1+1
37. 10=2+2+3+3	59. 10=3+3+1+3	81. 10=6+1+1+2
38. 10=2+2+4+2	60. 10=3+3+2+2	82. 10=6+1+2+1
39. 10=2+2+5+1	61. 10=3+3+3+1	83. 10=6+2+1+1
40. 10=2+3+1+4	62. 10=3+4+1+2	84. 10=7+1+1+1

Рассмотрим еще несколько частных случаев, зафиксировав $s = 4$ и уменьшая m от 10 до 4. Таблицы стандартных форм в этом случае имеют следующий вид.

Таблица 3. $m = 9$, $D_5^4 = 6$

1. 9=1+1+1+6
2. 9=1+1+2+5
3. 9=1+1+3+4
4. 9=1+2+2+4
5. 9=1+2+3+3
6. 9=2+2+2+3

Таблица 4. $m = 8$, $D_4^4 = 5$

1. 8=1+1+1+5
2. 8=1+1+2+4
3. 8=1+1+3+3
4. 8=1+2+2+3
5. 8=2+2+2+2

Таблица 5. $m = 7$, $D_4^3 = D_3^3 = 3$

1. 7=1+1+1+4
2. 7=1+1+2+3
3. 7=1+2+2+2

Таблица 6. $m = 6$, $D_2^1 = D_2^2 = 2$

1. 6=1+1+1+3
2. 6=1+1+2+2

Таблица 7. $m = 5$, $D_4^3 = D_1^4 = 1$

1. 5=1+1+1+2

Таблица 8. $m = 4$, $D_0^4 = 1$

1. 4=1+1+1+1

Объединим информацию, содержащуюся в таблицах 1–8, в следующую таблицу.

Таблица 9. Сравнение длин кодов поля кратности при кодировании двумя и одним числом для $m = 10, 9, 8, 7, 6, 5, 4$ и $s = 4$.

m	D_{m-s}^s	$\{\log_2 D_{m-s}^s\}$	$\sigma_{\max}$	$\{\log_2 \sigma_{\max}\}$	C_{m-1}^{s-1}	$\mathcal{L}_1(m,s)$	$\mathcal{L}_2(m,s)$
10	9	4	24	5	84	9	7
9	6	3	12	4	56	7	6
8	5	3	12	4	28	7	5
7	3	2	12	4	20	6	5
6	2	1	6	3	10	4	4
5	1	1	4	2	4	2	2
4	1	1	1	1	1	2	1

Результаты в последних двух столбцах табл. 9 подтверждают полученный выше из рассмотрения асимптотического поведения величин $\mathcal{L}_1(m,s)$ и $\mathcal{L}_2(m,s)$ вывод о том, что кодирование поля кратности одним числом более экономно.

Список использованной литературы

1. *Толстопятов А. А.* О возможности использования булевых уравнений для сжатия файлов // Вестник ИвГУ. – 2003. – Вып. 3. – С. 82–84.
2. *Толстопятов А. А.* Вычисление длины поля кратности при булевом сжатии файлов // Вестник ИвГУ. – 2004. – Вып. 3. – С. 71–76.

3. *Толстопятов А. А.* Быстрый алгоритм кодирования и декодирования поля порядка при булевом сжатии файлов // Математика и ее приложения: Журн. Иванов. матем. об-ва. – 2007. – Вып. 1 (4). – С. 35–46.

4. *Толстопятов А. А.* Медленный алгоритм кодирования и декодирования поля кратности при булевом сжатии файлов // Математика и ее приложения: Журн. Иванов. матем. об-ва. – 2007. – Вып. 1 (4). – С. 47–52.

5. *Толстопятов А. А.* Быстрый алгоритм кодирования и декодирования поля кратности при булевом сжатии файлов // Математика и ее приложения: Журн. Иванов. матем. об-ва. – 2007. – Вып. 1 (4). – С. 53–78.

6. *Толстопятов А. А., Хашин С. И.* Алгоритм построения поля порядка при булевом сжатии // Вестник ИвГУ. – 2004. – Вып. 3. – С. 139–143.

Поступила в редакцию 29.12.2008.