Математика и её приложения: ЖИМО. 2009. Вып. 1 (6). С. 121 – 128.

УДК 512.54

А. А. Толстопятов[1]

Возможность кодирования поля кратности и поля порядка одним числом

Ключевые слова: булево сжатие, поле кратности, поле порядка.

Рассматривается возможность объединения поля кратности и поля порядка в одно поле. Сравниваются длины кодов при кодировании этих полей как двумя числами, так и, как одно поле, — одним числом.

We consider the possibility of combining the multiplicity field and the order field in one field. We compare the code lengths for encoding of these fields both by two numbers and by one number (as for one field).

1. Введение

При булевом сжатии файлов код отдельного буфера содержит три поля: поле принадлежности, поле кратности, поле порядка и еще одно общее для всех буферов поле. В [1] был предложен быстрый алгоритм кодирования поля кратности, причем код содержал два числа — номер типа повторов в таблице стандартных форм всех возможных типов повторов и номер в таблице всех перестановок для данной стандартной формы. Алгоритм кодирования достаточно сложен. В [2] удалось объединить таблицу стандартных форм и таблицы перестановок внутри каждой стандартной формы. В результате получился быстрый алгоритм кодирования поля кратности. Этот алгоритм не только давал меньшую длину кода поля порядка, чем при кодировании этого поля двумя числами, но и был существенно проще последнего. Для такой объединенной таблицы поля кратности удалось построить и медленный алгоритм кодирования, когда таблица проходится последовательно строка за строкой, а не целыми блоками, как при быстром алгоритме [3]. Поэтому возникает идея об объединении в одно поле полей кратности и порядка. Но прежде, чем строить соответствующие алгоритмы, имеет смысл оценить, как такое объединение отразится на суммарной длине кодов полей кратности и порядка.

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

2. Постановка задачи

Пусть буфер содержит m кортежей, каждый длиной в n битов. Пусть s из них — разные. Пусть $n_1, n_2, \ldots, n_s$ — числа повторов 1-го, 2-го, …, s-го буферов. Тогда

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

Если поле кратности, т. е. поле, код которого позволяет восстанавливать числа повторов n_k, кодировать одним числом, когда наборы $n_1, n_2, \ldots, n_s$ рассматриваются как упорядоченные и считаются разными, если этот порядок для них разный, хотя сами числа повторов могут совпадать, то таблица всех таких наборов, удовлетворяющих (1), будет содержать C_{m-1}^{s-1} строк. Это значит, что длина кода поля кратности N_1 будет равна

$$N_1 = \{\log_2 C_{m-1}^{s-1}\} \geqslant \log_2 C_{m-1}^{s-1} \ . \tag{2}$$

Таблица поля кратности генерируется по набору чисел повторов $n_1, n_2, \ldots, n_s$ и является разной для разных стандартных форм, т. е. наборов чисел повторов, факторизованных по порядку внутри этих наборов. Такая таблица содержит

$$m! \left(\prod_{k=1}^{s} n_k! \right)^{-1}$$

строк. Поэтому длина поля кода порядка N_2 будет равна:

$$N_2 = \left\{ \log_2 \frac{m!}{\prod\limits_{k=1}^{s} n_k!} \right\} > \log_2 \frac{m!}{\prod\limits_{k=1}^{s} n_k!} \ . \tag{3}$$

Однако, оценить суммарную длину кода полей кратности и порядка просто сложив N_1 из (2) и N_2 из (3) было бы неверным, т. к. на наборы чисел повторов, стоящих в конце таблицы могут приходится не самые длинные, а наоборот, самые короткие таблицы перестановок. Проще всего в этом убедиться, рассмотрев малоразмерный пример. Возьмем $m = 6;\ s = 3$. Тогда таблица поля кратности содержит $C_{m-1}^{s-1} = C_5^2 = 10$ строк. Значит $N_1 = \{\log_2 10\} = 4$. Самая длинная таблица поля порядка будет получаться при $n_1 = n_2 = n_3 = 2$. Тогда она будет содержать

$$\frac{m!}{n_1! n_2! n_3!} = \frac{6!}{2!^3} = 90$$

строк. Но тогда из (3) найдем, что $N_2 = \{\log_2 90\} = 7$. Если бы мы считали, что суммарная длина кодов полей кратности и порядка есть

$$N = N_1 + N_2, \tag{4}$$

то у нас получилось бы, что $N = 7 + 4 = 11$. Если обозначить через σ число строк в таблице поля кратности, т. е. положить

$$\sigma = \frac{m!}{\prod\limits_{k=1}^{s} n_k!}, \tag{5}$$

то для рассматриваемого случая можно составить следующую таблицу.

Таблица 1. Суммарная длина кодов полей кратности и порядка

№	Тип повторов	$N_1 = \{\log_2 N\}$	σ	$N_2 = \{\log_2 \sigma\}$	$N_1 + N_2$
1	1+1+4	1	30	5	6
2	1+2+3	1	60	6	7
3	1+3+2	2	60	6	8
4	1+4+1	2	30	5	7
5	2+1+3	3	30	5	9
6	2+2+2	3	90	7	10
7	2+3+1	3	60	6	9
8	3+1+2	3	60	6	9
9	3+2+1	4	60	6	10
10	4+1+1	4	30	5	9

Таблица 1 показывает, что максимальная суммарная длина кодов полей кратности и порядка равна 10 и меньше 11. Поэтому оценку суммарной длины кодов полей кратности и порядка надо определять по максимальному числу в последнем столбце таблицы типа табл. 1, но построенной для произвольных m и s. Значит, первая задача, которая возникает, — найти этот максимум. Заметим, что таблицу типа табл. 1 можно организовать по-разному. Таблица 1 устроена так, что сначала идут типы повторов, для которых $n_1 = 1$, потом $n_1 = 2$ и т. д. и такое же упорядочение выполняется для n_2, n_3, …, n_s. Но можно организовать эту таблицу по-другому, используя понятие стандартной формы. Для данных m и s разных стандартных форм будет D^s_{m-s}, где D^s_n может быть задана следующими рекуррентными формулами

$$\begin{aligned}
&1.\ D^s_n &&= D^s_{n-s} + D^{s-1}_n, && s \leqslant n,\\
&2.\ D^s_n &&= D^n_n, && s > n,\\
&3.\ D^s_0 &&= 1,\\
&4.\ D^1_n &&= 1,\\
&5.\ D^2_n &&= \left[\tfrac{n}{2}\right] + 1.
\end{aligned} \tag{6}$$

Для всех типов повторов, принадлежащих одной стандартной форме таблица поля порядка одна и та же. Поэтому таблицу типа табл. 1 можно устроить так, что сначала будут идти типы повторов, принадлежащих

первой стандартной форме, потом второй и т. д. Сами же стандартные формы можно упорядочить лексикографически. Для нашего примера с $m = 6;\ s = 3$, число стандартных форм будет $D^s_{m-s} = D^3_3$. С помощью (6), найдем $D^3_3 = 3$. Лексикографически упорядоченные стандартные формы дают таблицу:

$$\begin{array}{ll} 1. & 6 = 1 + 1 + 4, \\ 2. & 6 = 1 + 2 + 3, \\ 3. & 6 = 2 + 2 + 2. \end{array} \tag{7}$$

Увеличивая каждую строчку таблицы (7) до целой таблицы, включающей в себя все возможные перестановки чисел повторов, принадлежащих данной стандартной форме, получим вместо табл. 1 следующую таблицу.

Таблица 2. Суммарная длина кодов полей кратности и порядка для типов повторов, упорядоченных с помощью стандартных форм

№	Тип повторов	$N_1 = \{\log_2 N\}$	σ	$N_2 = \{\log_2 \sigma\}$	$N_1 + N_2$
1	1+1+4	1	30	5	6
2	1+4+1	1	30	5	6
3	4+1+1	2	30	5	7
4	1+2+3	2	60	6	8
5	1+3+2	3	60	6	9
6	2+1+3	3	30	6	9
7	2+3+1	3	60	6	9
8	3+1+2	3	60	6	9
9	3+2+1	4	60	6	10
10	2+2+2	4	90	7	11

Сравнение таблиц 1 и 2 показывает на этом частном примере, что максимальная суммарная длина кодов кратности и порядка зависит от способа упорядочения таблицы типов повторов при фиксированных m и s. Разумеется, способы упорядочения типов повторов, реализованные в таблицах 1 и 2 не единственные. Но, в определенном смысле, они являются универсальными. Поэтому только ими мы и ограничимся.

Если через $\mathcal{N}_1$ обозначить $\max(N_1 + N_2)$, а через $\mathcal{N}_2 = \{\log_2 \eta\}$, где η — число строк в таблице объединенных полей кратности и порядка, то задача о целесообразности кодирования полей кратности и порядка одним числом сводится к вопросу о выполнении неравенства

$$\mathcal{N}_2 < \mathcal{N}_1, \tag{8}$$

и тогда такое кодирование имеет смысл, или — противоположного неравенства

$$\mathcal{N}_2 > \mathcal{N}_1, \tag{9}$$

и тогда такое кодирование ничего не дает.

3. Максимальное значение σ

Для нахождения $\mathcal{N}_1 = \max(N_1 + N_2)$ необходимо найти максимальное значение σ. Из (5) видно, что σ будет максимально тогда, когда минимальным будет $\prod_{k=1}^{s} n_k!$ при фиксированных m и s и выполнены условия (1). Для нахождения стандартной формы с минимальными $\prod_{k=1}^{s} n_k!$ заметим, что существует единственная стандартная форма, в которой все числа повторов или равны между собой или отличаются на единицу. Для этой стандартной формы числа повторов есть:

$$\begin{array}{ll} n_k & = \left[\frac{m}{s}\right], \quad k = 1, 2, \ldots, p, \\ n_{p+l} & = \left[\frac{m}{s}\right] + 1, \quad l = 1, 2, \ldots, s - p, \end{array} \tag{10}$$

где

$$p = \left(\left[\frac{m}{s}\right] + 1\right) s - m \ . \tag{11}$$

Все остальные стандартные формы могут быть получены из (10) последовательным увеличением одного из n_k на 1 и увеличением другого тоже на 1. Разница между разными n_k в (10) минимальна: она равна 0 или 1. Нетрудно убедиться, что при увеличении этой разницы между двумя множителями $a!b!$ это произведение возрастает. Действительно, пусть $b > a$, а значит $b = a + k$. Сравним $a!(a + k)!$ и $(a - 1)!(a + k + 1)!$. Предположим, что

$$a!(a + k)! < (a - 1)!(a + k + 1)! \ . \tag{12}$$

Тогда, записав (12) в виде $a(a - 1)!(a + k)! < (a - 1)!(a + k + 1)(a + k)!$ и сократив одинаковые множители в этом неравенстве, приходим к верному неравенству $a < a + k + 1$, а значит и (12) справедливо. Это означает, что, поскольку в любой стандартной форме в произведении $\prod_{k=1}^{s} n_k!$ в отличие от (10) хотя бы у одной пары множителей разница между числами, от которых берутся факториалы, больше, чем у любой аналогичной пары в (10), то и $\prod_{k=1}^{s} n_k!$ для такой стандартной формы будет больше, чем для (10). Поэтому максимальное значение σ для фиксированных m и s есть

$$\sigma_{\max} = \frac{m!}{\left(\left[\frac{m}{s}\right]!\right)^{p} \left[\left(\left[\frac{m}{s}\right] + 1\right)!\right]^{s-p}} \ , \tag{13}$$

где p дается формулой (11).

Нетрудно найти не только $\sigma_{\max}$, но и $\sigma_{\min}$. Действительно, стандартная форма, в которой перемещение единицы от одного числа повторов к

другому будет только уменьшать разницу между этими числами, есть

$$\begin{aligned} &n_k = 1, \quad k = 1, 2, \ldots, s-1, \\ &n_s = m - s + 1. \end{aligned} \tag{14}$$

Поэтому из (5) и (14) следует, что

$$\sigma_{\min} = \frac{m!}{(n-s+1)!}\,. \tag{15}$$

4. Длина кода полей кратности и порядка при отдельном их кодировании

Оценка максимальной длины совместного кода полей кратности и порядка зависит от способа упорядочения наборов чисел повторов. Для этой оценки надо вычислить номер типа повторов в этой таблице, для которого σ достигает максимума. Таких номеров будет ξ, где

$$\xi = \frac{s!}{p!(s-p)!} = C_s^p\,. \tag{16}$$

Нам нужно найти самый большой из этих ξ номеров. Этот номер есть номер следующего типа повторов:

$$\begin{aligned} &n_1 = n_2 = \cdots = n_{s-p} = \left[\tfrac{m}{s}\right] + 1\ , \\ &n_{s-p+1} = n_{s-p+2} = \cdots = n + s = \left[\tfrac{m}{s}\right]\ . \end{aligned} \tag{17}$$

Согласно алгоритму, построенному в [2], для (17) искомый номер θ можно найти согласно формуле

$$\theta = \sum_{j=1}^{p-s}\sum_{k=2}^{\left[\frac{m}{s}\right]+1} C^{s-j-1}_{m-\left(\left[\frac{m}{s}\right]+1\right)(j-1)-k} + \sum_{j=1}^{p}\sum_{k=2}^{\left[\frac{m}{s}\right]} C^{s-j-1}_{m-\left[\frac{m}{s}\right](j-1)-k}. \tag{18}$$

Тогда, учитывая, что $N_1 = \{\log_2\theta\}$, (13) и $N_2 = \{\log_2\sigma_{\max}\}$, найдем для $\mathcal{N}_1$ следующее выражение:

$$\mathcal{N}_1 = \{\log_2\theta\} + \{\log_2\sigma_{\max}\}, \tag{19}$$

где p определяется из (11).

В случае упорядочения таблицы всех типов повторов по аналогии с таблицей 2, тип повторов (17) всегда будет стоять на C_{m-1}^{s-1}-м месте в этой таблице, а значит

$$\theta = C_{m-1}^{s-1}, \tag{20}$$

и в этом случае

$$\mathcal{N}_1 = \{\log_2 C_{m-1}^{s-1}\} + \{\log_2\sigma_{\max}\}. \tag{21}$$

5. Сравнение длин кодов полей кратности и порядка при кодировании их одним или двумя числами

В случае объединения полей кратности и порядка в одно поле, каждая строчка таблицы стандартных форм превращается в блок таблицы объединенных полей, имеющей длину, равную соответствующей σ; разные упорядочения таблицы, разумеется, не влияют на ее длину, т. е. число строк в этой таблице η которое равно

$$\eta = m! \sum_{i=1}^{C_{m-1}^{s-1}} \frac{1}{\prod\limits_{k=1}^{s} n_k^i!}, \tag{22}$$

где через n_k^i обозначено k-е число повторов в i-м типе повторов. Формула для η может быть упрощена в том смысле, что сумма будет содержать меньше членов, если под $\tilde{n}_k^i$ понимать k-е число повторов в i-ой стандартной форме, когда все типы повторов объединены в классические формы, в которых одинаковые числа повторов, но с разным их порядком, считаются неразличающимися. Тогда вместо (22) будем иметь

$$\eta = m! \sum_{i=1}^{D_{m-s}^{s}} \frac{1}{\prod\limits_{k=1}^{s} \tilde{n}_k^i!}. \tag{23}$$

Из (22) или (23) будем иметь для $\mathcal{N}_2$ следующее выражение:

$$\mathcal{N}_2 = \left\{ \log_2 m! \sum_{i=1}^{C_{m-1}^{s-1}} \frac{1}{\prod\limits_{k=1}^{s} \tilde{n}_k^i!} \right\} = \left\{ \log_2 m! \sum_{i=1}^{D_{m-s}^{s}} \frac{1}{\prod\limits_{k=1}^{s} \tilde{n}_k^i!} \right\}. \tag{24}$$

Подставляя (19) и (24) в неравенства (8) и (9), получаем критерий, допускающий оценку лучшего кодирования в смысле меньшей длины кода полей кратности и порядка, если таблица типов повторов упорядочена аналогично табл. 1.

Если выполнено верхнее неравенство, то более экономичный код одним числом, если нижнее — то двумя.

Аналогично, заменяя в (24) формулу (19) на (21), получим критерий для упорядочения таблицы типов повторов аналогичному табл. 2.

Наконец, отметим, что обнаруженная зависимость максимальной длины кодов полей кратности и порядка от способа упорядочивания таблицы типа повторов, позволяет поставить задачу о нахождении оптимального

упорядочивания. Поскольку эта оптимизация имеет смысл не для отдельного разбиения файла и не для отдельного файла, а для их класса, такого как текст, графика, exe и т. д., то уже это делает целесообразным построение алгоритмов кодирования двух полей одним числом.

Список литературы

1. *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.
2. Толстопятов А. А., Гришко М. Е. Быстрый алгоритм кодирования и декодирования поля кратности одним числом при булевом сжатии файлов // Вестник ИвГУ. – 2009. – Вып. 2. – С. 45–52.
3. Толстопятов А. А., Гришко М. Е. Медленный алгоритм кодирования и декодирования поля кратности одним числом при булевом сжатии файлов // Вестник ИвГУ. – 2009. – Вып. 2. – С. 53–55.

Поступила в редакцию 27.12.2009