Математика и её приложения: ЖИМО. 2007. Выпуск 1 (4). С. 79 – 92.

УДК 512.54

А. А. Толстопятов[1]

Факторы адаптации при булевом сжатии файлов

Ключевые слова: булева алгебра, сжатие информации.

Рассматривается структура кода при булевом сжатии файлов. Сформулирован критерий сжатия. Обсуждены факторы адаптации кода, подбор которых позволяет удовлетворить критерию сжатия и сократить длину кода отдельных полей.

We consider the structure of a Boolean compression code, and formulate the criterion of file-compression. Also we discuss those factors of adaptation of a code, which provide as to satisfy of a compression criterion, and to shorten the code length of individual fields.

Целью настоящей работы является подведение текущих итогов исследований по булеву сжатию файлов и обсуждение дальнейших задач в этом направлении. На настоящем этапе исследований было показано, что обойти теорему Шеннона, ограничивающую возможность сжатия файлов условием существования в разбиении файла на кортежи повторяющихся кортежей, можно, если самому разбиению придана определенная структура [1]. А именно, если файлы разбить на кортежи длиной n битов, а потом объединить их в L буферов с длинами nm_l, $l = 1, 2, \ldots, L$, где m_l — число кортежей в l-м буфере и рассматривать кортежи, входящие в отдельный буфер, как решение булева уравнения от n переменных $x_1, x_2, \ldots, x_n$, то код отдельного буфера будет состоять из трех полей.

1. *Поле принадлежности.* Оно задается коэффициентами булева уравнения и имеет длину L_1, равную

$$L_1 = 2^n. \tag{1}$$

2. *Поле кратности.* Оно задается номером в лексикографически упорядоченой таблице всех чисел повторов $n_1, n_2, \ldots, n_s$, где s_l — число разных кортежей, входящих в l-й буфер. Это поле имеет длину L_2, равную

$$L_2 = \log_2 D^s_{m-s}. \tag{2}$$

Определение величины D^s_{m-s} дано в [3].

3. *Поле порядка.* Оно задается номером в лексикографически упорядоченной таблице перестановок кортежей, входящих в данный буфер, с учетом кратности кортежей, и имеет длину L_3, равную

$$L_3 = \log_2 \frac{m!}{\prod\limits_{k=1}^{s} n_k!}. \tag{3}$$

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

Необходимость двух последних полей вызвана тем, что булева переменная $x \in GF(2)$, удовлетворяет условиям

$$x + x = 0, \quad x^2 = x, \tag{4}$$

а поэтому код поля принадлежности позволяет определить кортежи, входящие в данный буфер, во-первых, без учета их кратности, и, во-вторых, в порядке, вообще говоря, завищящем от способа решения булева уравнения. Коэффициент сжатия отдельного буфера есть

$$k = \frac{mn}{L_1 + L_2 + L_3} = \frac{mn}{2^n + \log_2 D^s_{m-s} + \log_2 \frac{m!}{\prod\limits_{k=1}^{s} n_k!}}. \tag{5}$$

Очевидно, для сжатия файла вовсе не нужно, чтобы $k > 1$, поэтому отдельные буферы могут растягиваться, а сжатие файла будет происходить за счет большего сжатия других буферов. Коэффициент сжатия всего файла, если величины, относящиеся к буферу снабдить индексом $l = 1, 2, \ldots, L$, будет равен

$$k = \frac{n \sum\limits_{l=1}^{L} m_l}{2^n L + \sum\limits_{l=1}^{L} \log_2 D^{s_l}_{m_l - s_l} + \sum\limits_{l=1}^{L} \log_2 \frac{m_l!}{\prod\limits_{k=1}^{s} n_k!}}, \tag{6}$$

и условие сжатия есть

$$k > 1, \tag{7}$$

где k задано формулой (6). Нетрудно убедиться, что выполнение условия (7) накладывает ограничения на значение m_l. Проще всего это сделать, рассмотрев отдельный буфер, когда нет повторов, т. е.

$$s = m, \; n_1 = n_2 = \ldots = n_m = 1.$$

Тогда (5) принимает вид

$$k = \frac{mn}{2^n + \log_2 m!}, \tag{8}$$

или в результате использования формулы Стирлинга $m! \approx \left(\frac{m}{e}\right)^m$ следующий вид:

$$k \approx \frac{mn}{2^n + m(\log_2 m - \log_2 e)}. \tag{9}$$

Будем искать m в виде

$$m = 2^n \varepsilon, \tag{10}$$

т. к. при условии $m = s,\ m = 1, 2, \ldots, 2^n$. Из (9) и (10) получим:

$$k \approx \frac{1}{1 + \frac{1}{n} F(\varepsilon)}, \tag{11}$$

где

$$F(\varepsilon) = \frac{1}{\varepsilon} + \log_2 \varepsilon - \log_2 e. \tag{12}$$

Поскольку $\lim\limits_{\varepsilon \to 0} F(\varepsilon) = +\infty$, $\lim\limits_{\varepsilon \to \infty} F(\varepsilon) = +\infty$ и функция $F(\varepsilon)$ имеет единственный минимум при $\varepsilon = \ln 2$, равный

$$F_{\min} = F(\ln 2) = -\log_2 \log_2 e \approx -0.4, \tag{13}$$

то существуют два положительных корня ε_1 и ε_2 уравнения

$$F(\varepsilon) = 0, \tag{14}$$

причем в этом случае $k = 1$. А значит, если $\varepsilon_1 < \varepsilon < \varepsilon_2$, то $k > 1$. Максимальное значение коэффициента сжатия $k_{\max}$ будет равно

$$k_{\max} = \frac{1}{1 - \frac{0.4}{n}}. \tag{15}$$

Корни ε_1 и ε_2 уравнения (14), т. е. границы $m_1 = 2^n \varepsilon_1$ и $m_2 = 2^n \varepsilon_2$, попадая в которые число кортежей m $(m_1 < m < m_2)$ в буфере обеспечивает $k > 1$, нетрудно найти. Действительно, т. к. $F(1/4) = 2 - \log_2 e > 0$, а $F(1/2) = 1 - \log_2 e < 0$, то $1/4 < \varepsilon_1 < 1/2$. Продолжая деление интервала $(1/4,\ 1/2)$ пополам так, чтобы функция $F(\varepsilon)$ на концах интервала принимала значения разных знаков, до тех пор, пока не получится обращение $F(\varepsilon)$ в нуль с требуемой точностью, находим с точностью до 10^{-3}, что

$$\varepsilon_1 \approx 0.324. \tag{16}$$

Аналогично для ε_2, начиная с интервала $(1, 2)$, т. к. $F(1) = 1 - \log_2 e < 0$, а $F(2) = 3/2 - \log_2 e > 0$, с той же точностью найдем

$$\varepsilon_2 \approx 1.906. \tag{17}$$

Это значит, что

$$m_1 \approx 0.324 \cdot 2^n, \quad m_2 \approx 1.906 \cdot 2^n, \quad m_{\max} \approx 2^n \ln 2 \approx 0.69 \cdot 2^n.$$

Полученные результаты показывают, что в этом случае никакого сжатия не происходит. Это можно проиллюстрировать, составив таблицу для $n = 8,\ 16$.

Таблица 1. Оценка параметров булева сжатия

n	2^n	m_1	$m_{\max}$	m_2	$k_{\max}$
8	256	83	177	488	1.05
16	65 536	21.234	45.220	124.928	1.03

Полученные оценки в общем подтверждаются результатами тестирования программы, написанной для реализации описанной выше структуры кода. При выборе параметров вместо (9) была использована более грубая формула

$$k = \frac{mn}{2^n + m\log_2 m}. \tag{18}$$

Так как $mn > 2^n$, то $m > \frac{2^n}{n} = m$, и $mn > m\log_2 m \Longrightarrow m < 2^n = m_2$. Взяв в качестве m число

$$m = \frac{1}{2}(m_1 + m_2) = \frac{2^{n-1}(n+1)}{n}$$

и выбрав $n = 8$, для которого $m = 144$, для различных типов файлов, а именно, текстовых, графических, аудио, exe и т. д., можно вычислить коэффициент сжатия по точной формуле (6). Тестирование показало, что $L_1 = 32$ байта, $L_2 = 4$ байта, $L_3 = 104$ байта, а коэффициент сжатия равен $k = 1 \pm 0.07$, т. е. действительно происходит перекодировка. Но в результате этой перекодировки код файла приобретает определенную структуру, используя которую можно пытаться уменьшить L_1, L_2 или L_3.

1. Структура кода и условия сжатия

Поскольку код состоит из трех полей, то увеличить коэффициент сжатия можно, уменьшая длину кода любого из этих трех полей. Для этого код нужно адаптировать к конкретному файлу, подбирая параметры кода, которые будем называть факторами адаптации. Подбор таких факторов, должен приводить к выполнению одного или нескольких условий для длин кода трех полей. А именно, для поля принадлежности должно выполняться условие

$$L_1 < 2^n. \tag{19}$$

При этом факторами адаптации служат число буферов L и число кортежей в буферах m_l, $l = 1, 2, \ldots, L$.

Для поля кратности должно выполняться условие

$$L_2 < \log_2 D^s_{m-s}. \tag{20}$$

При этом фактором адаптации, помимо тех, что указаны для поля принадлежности, является способ упорядочения таблицы чисел повторов [3, 5, 6], такой, чтобы как можно большее число буферов имели как можно

меньшие номера. Однако, т. к. длина кода поля кратности является самой маленькой среди длин кодов трех полей, и кроме того, при выполнении условия $m \ll 2^n$ почти все числа кратности равны единице, то вряд ли за счет сокращения длины кода этого поля удастся сколько-нибудь существенно увеличить коэффициент сжатия.

Для поля порядка должно выполняться условие

$$L_3 < \log_2 \frac{m!}{\prod\limits_{k=1}^{s} n_k!}. \tag{21}$$

Факторами адаптации для кода этого поля, помимо L и m_l, $l = 1, 2, \ldots, L$, является способ упорядочения таблицы перестановок [4, 7] и способ решения булева уравнения из поля принадлежности, т. к. порядок, в котором восстанавливаются кортежи, зависит от этого способа. Поскольку при тех m, которые обеспечивают сжатие, длина кода поля кратности самая большая, то за счет этого фактора коэффициент сжатия можно существенно увеличить. Однако, разных способов решения булева уравнения гораздо меньше, чем способов разместить в разном порядке одни и те же кортежи в буфере, что ограничивает возможность увеличения коэффициента сжатия за счет адаптации кода поля порядка.

Основным способом адаптации кода является способ, приводящий к выполнению условия (7). Это связано с тем, что все кортежи, на которые разбивается буфер, содержатся в линейном векторном пространстве над полем $GF(2)$, что открывает возможность поиска линейных зависимостей между кортежами, входящими в каждый буфер, и рассматриваемых как элементы такого пространства. Однако, тестирование программы поиска таких линейных зависимостей на реальных файлах показало, что этих зависимостей нет. Если код поля принадлежности является коэффициентами булева уравнения, то поскольку булевы полиномы можно не только умножать на элементы $GF(2)$ и складывать, но и умножать друг на друга, т. е. булевы полиномы образуют алгебру, то это открывает возможность поиска нелинейных зависимостей между булевыми полиномами, коэффициенты которых кодируют поле принадлежности. Но именно потому, что эти зависимости существуют между кодами полей принадлежности всех буферов, условию (7) нельзя удовлетворить для отдельного буфера, а только для всего файла. Эта адаптация является характерной и специфической именно для булева сжатия, поэтому ей будет уделено далее больше всего внимания.

2. Код поля принадлежности

Построение кода поля принадлежности и алгоритм, приводящий к выполнению условия (7), рассмотрены в работе [2]. Однако, в этой работе не была рассмотрена одна возможность, которую обсудим ниже. А именно, пусть f_l, $l = 1, 2, \ldots, L$, — булевы полиномы полей принадлежности всех буферов. Полиномы $f_l(x_1, x_2, \ldots, x_n)$, где x_j, $j = 1, 2, \ldots, n$,

есть булевы переменные, удобно рассматривать не в базисе Жегалкина $1, x_j, x_ix_j, \ldots, x_1x_2\ldots x_n$, а в базисе Лагранжа $L_j(x_i)$, т. е. задать коэффициентами C_j^l, $j = 0, 1, \ldots, 2^n - 1$, $l = 1, 2, \ldots, L$, в разложении

$$f_l = \sum_{j=0}^{2^n-1} C_j^l L_j. \tag{22}$$

Тогда, если $\varphi_p(x_i)$, $p = 1, 2, \ldots, P$, — порождающие булевы полиномы, через которые можно выразить все f_l с помощью кодирующего полинома $F_P(e_i^l)$ — булева полинома степени P от булевых полиномов e_i^l, $l = 1, 2, \ldots, L$, $i = 1, 2, \ldots, I$, в качестве которых можно брать любой из φ_P, причем

$$F_P(e_i^l) = a_0 + a_{i_1}e_{i_1}^l + a_{i_1i_2}e_{i_1}^l e_{i_2}^l + \ldots + a_{i_1i_2\ldots i_P}e_{i_1}^l e_{i_2}^l \ldots e_{i_p}^l, \tag{23}$$

и если разложить e_i^l по φ_P, а φ_P и F_P по базису Лагранжа,

$$e_i^l = \sum_{p=1}^{P} C_{ip}^l \varphi_p \ , \tag{24}$$

$$\varphi_p = \sum_{j=0}^{2^n-1} C_{pj} L_j \ , \tag{25}$$

$$F_P = \sum_{j=0}^{2^n-1} F_{Pj} L_j \ , \tag{26}$$

то

$$F_P = a_0 + a_{i_1} C_{i_1p_1}^l C_{p_1j} + a_{i_1i2} C_{i_1p_1}^l C_{i_2p_2}^l C_{p_1j} C_{p_2j} + \cdots + \\ + a_{i_1i2\ldots i_P} C_{i_1p_1}^l C_{i_2p_2}^l \ldots C_{i_Pp_P}^l C_{p_1j} C_{p_2j} \ldots C_{p_Pj} = C_j^l. \tag{27}$$

Система булевых уравнений (27) определяет по заданным C_j^l, т. е. по заданным f_l, не только коэффициенты $a_0, \ldots, a_{i_1i_2\ldots i_P}$ кодирующего полинома F_P, но и порождающие φ_P, заданные коэффициентами C_{pj}, и кортежи e_i^l, построенные из φ_P, заданные коэффициентами C_{ip}^l. Если уравнения (27) имеют хотя бы одно решение, то между f_l существует, вообще говоря, нелинейные зависимости, и длина кода поля принадлежности L_1 будет удовлетворять условию (19), если выполнено

$$L_1 = I \log_2 P < 2^n. \tag{28}$$

При этом 2^P коэффициентов кодирующего полинома F_P в код не включались. Это требовало изучения устойчивости полинома F_P при изменении файла, которое оказалось исключительно сложным и не было доведено до конца. Однако, условие (28) можно изменить, включив в код коэффициенты кодирующего полинома F_P, что, конечно, уменьшит возможность

подбора таких адаптирующих факторов, как P и I, но зато избавит от требования устойчивости F_P на том или ином классе файлов. Тогда вместо (28) будем иметь условие

$$L_1 = 2^P/L + I\log_2 P < 2^n. \tag{29}$$

Поэтому если выполнено (28) параметру адаптации P нужно последовательно придавать значения $P = 2, 3, \ldots, 2^{2^n} - 1$, то условие (28) — это ограничение на допустимые значения I:

$$I < \frac{2^n}{\log_2 P}. \tag{30}$$

Аналогично и в случае (28), $P = 2, 3, \ldots, n-1$, а тогда для I имеем ограничение

$$I < \frac{2^n - 2^P/L}{\log_2 P}. \tag{31}$$

Критерием возможности сжатия файла с максимальным коэффициентом сжатия, таким образом, оказывается существование минимального P и минимального I, удовлетворяющих (30) или (31), при которых уравнение (27) имеет хотя бы одно решение. Однако, поскольку C_j^l в (27) зависит от таких параметров адаптации как L и m_l, $l = 1, 2, \ldots, L$, т. е. от разбиения файла, то центральной задачей возможности осуществления булева сжатия является задача о критерии существования разбиения файла и алгоритма построения такого разбиения, для которого уравнения (27) при ограничении (30) имеют хотя бы одно решение.

Рассмотрим вопрос об оценке коэффициента сжатия при выполнении условия (18). Если в (9) заменить длину кода поля принадлежности $L_1 = 2^n$ на $L_1 = I\log_2 P$ или $L_1 = 2^P/L + I\log_2 P$ и задать m в виде $m = 2^n\varepsilon$, то вводя обозначение

$$\alpha = \frac{I\log_2 P}{2^n} \tag{32}$$

или

$$\alpha = \frac{2^P/L + I\log_2 P}{2^n}, \tag{33}$$

вместо (9) будем иметь

$$k = \frac{1}{1 + \frac{1}{n}F_\alpha(\varepsilon)}, \tag{34}$$

где

$$F_\alpha(\varepsilon) = \frac{\alpha}{\varepsilon} + \log_2\varepsilon - \log_2 e. \tag{35}$$

Функция $F_\alpha(\varepsilon)$ достигает минимального значения при $\varepsilon = \alpha \ln 2$, оно равно

$$F_{\alpha_{\min}} = F_\alpha(\alpha \ln 2) = -\log_2 \log_2 e + \log_2 \alpha. \tag{36}$$

Если выполняется (28) или (29), то

$$0 < \alpha < 1, \tag{37}$$

а значит $F_{\alpha_{\min}} < 0$ и, следовательно, $k > 1$, т. к.

$$k_{\max} = \frac{1}{1 - \frac{1}{n}(\log_2 \log_2 e - \log_2 \alpha)}. \tag{38}$$

Подставляя (32) в (38), найдем, что

$$k_{\max} = \frac{n}{\log_2(I \ln P)}. \tag{39}$$

Подставляя (33) в (38), найдем, что

$$k_{\max} = \frac{n}{P - \log_2 L + \log_2 \left(1 + \frac{I \log_2 P}{2^P}\right) - \log_2 \log_2 e}. \tag{40}$$

Пренебрегая в (40) медленно растущим логарифмическим членом и постоянной в знаменателе, находим нужную оценку для $k_{\max}$ в случае (33)

$$k_{\max} = \frac{n}{P - \log_2 L}, \tag{41}$$

которая показывает, что $k_{\max}$ увеличивается с ростом n, причем этот рост расширяет возможности кода $P = 2, 3, \ldots, n-1$ для выполнения условия (27).

Для определения границ m_1 и m_2 числа кортежей в буфере, внутри которых выполняется условие $k > 1$, нужно найти решение уравнения

$$F_\alpha(\varepsilon) = 0, \tag{42}$$

где $F_\alpha(\varepsilon)$ задано (35). Нетрудно увидеть, что вследствие (37) для любого ε выполнено неравенство

$$F_\alpha(\varepsilon) = F(\varepsilon) + \frac{\alpha - 1}{\varepsilon} < F(\varepsilon). \tag{43}$$

Из (43) следует, что график функции $F_\alpha(\varepsilon)$ целиком лежит под графиком функции $F(\varepsilon)$. А это значит, что уравнение (42) имеет два корня ε_1^α и ε_2^α, причем

$$\varepsilon_1^\alpha = \varepsilon_1 - \Delta_1 \ , \text{ и } \varepsilon_2^\alpha = \varepsilon_2 - \Delta_2 \ , \tag{44}$$

где Δ_1 и Δ_2 удовлетворяют условиям

$$0 < \Delta_1 < \varepsilon_1 \quad \text{и} \quad 0 < \Delta_2 < \varepsilon_2. \tag{45}$$

Так как

$$F(\varepsilon_1) = \frac{1}{\varepsilon_1} + \log_2 \varepsilon_1 - \log_2 e = 0, \tag{46}$$

то из (35) и (42) найдем, что

$$F_\alpha(\varepsilon_1^\alpha) = \frac{\alpha}{\varepsilon_1 - \Delta_1} - \frac{1}{\varepsilon_1} + \log_2 \left(1 - \frac{\Delta_1}{\varepsilon_1}\right) = 0. \tag{47}$$

Если логарифм в (47) разложить в степенной ряд по малой величине $\dfrac{\Delta_1}{\varepsilon_1}$, то получим

$$(1 - \varepsilon_1 \log_2 e)\frac{\Delta_1}{\varepsilon_1} + \varepsilon_1 \log_2 e \sum_{n=2}^{\infty} \frac{1}{n(n-1)} \left(\frac{\Delta_1}{\varepsilon_1}\right)^n = 1 - \alpha. \tag{48}$$

Поправка Δ_1 может быть найдена обращением ряда в (48) с любой степенью точности в виде

$$\Delta_1 = \varepsilon_1 \sum_{n=0}^{\infty} a_n (1 - \alpha)^n. \tag{49}$$

Ограничиваясь в разложении (49) первыми двумя членами, получим

$$\Delta_1 = \frac{\varepsilon_1(1 - \alpha)}{1 - \varepsilon_1 \log_2 e} \ . \tag{50}$$

Аналогично для Δ_2, пользуясь тем, что

$$F(\varepsilon_2) = \frac{1}{\varepsilon_2} + \log_2(1 + \frac{\Delta_2}{\varepsilon_2}) = 0, \tag{51}$$

и раскладывая логарифм в (51) в степенной ряд по малой величине $\dfrac{\Delta_2}{\varepsilon_2}$, получим

$$(\varepsilon_2 \log_2 e - 1)\frac{\Delta_2}{\varepsilon_2} + \varepsilon_2 \log_2 e \sum_{n=2}^{\infty} \frac{(-1)^n}{n(n-1)} \left(\frac{\Delta_2}{\varepsilon_2}\right)^n = 1 - \alpha. \tag{52}$$

Поправка Δ_2 определяется из (52) обращением ряда в виде

$$\Delta_2 = \varepsilon_2 \sum_{n=0}^{\infty} b_n (1 - \alpha)^n. \tag{53}$$

Ограничиваясь в разложении (53) первыми двумя членами, получим

$$\Delta_2 = \frac{\varepsilon_2(1-\alpha)}{\varepsilon_2 \log_2 e - 1}. \tag{54}$$

Тогда для m_1 и m_2 найдем из (10), (50), (54), что

$$m_1 = \frac{\varepsilon_1(1-\alpha)\cdot 2^n}{1-\varepsilon_1 \log_2 e}, \quad m_2 = \frac{\varepsilon_2(1-\alpha)\cdot 2^n}{\varepsilon_2 \log_2 e - 1}, \tag{55}$$

где ε_1 и ε_2 определяются (16) и (17). Подставляя (16), (17) и (32) в (55), находим, что

$$\begin{aligned} m_1 &= \frac{\varepsilon_1(1-\alpha)\cdot 2^n}{1-\varepsilon_1 \log_2 e} &&= 0.6(2^n - I\log_2 P)\ , \\ m_2 &= \frac{\varepsilon_2(1-\alpha)\cdot 2^n}{\varepsilon_2 \log_2 e - 1} &&= 1.08(2^n - I\log_2 P). \end{aligned} \tag{56}$$

Это значит, что при выполнении условия

$$0.6(2^n - I\log_2 P) < m < 1.08(2^n - I\log_2 P) \tag{57}$$

будут выполнены условия (7). Подставляя (16), (17) и (33) в (55), найдем, что

$$\begin{aligned} m_1 &= \frac{\varepsilon_1(2^n - 2^P/L - I\log_2 P)}{1-\varepsilon_1 \log_2 e} &&= 0.6(2^n - 2^P/L - I\log_2 P)\ , \\ m_2 &= \frac{\varepsilon_2(2^n - 2^P/L - I\log_2 P}{\varepsilon_2 \log_2 e - 1} &&= 1.08(2^n - 2^P/L - I\log_2 P). \end{aligned} \tag{58}$$

Это значит, что при выполнении условия

$$0.6(2^n - 2^P/L - I\log_2 P) < m < 1.08(2^n - 2^P/L - I\log_2 P) \tag{59}$$

будут выполнены условия (7).

Еще раз подчеркнем, что для сжатия файла вовсе не обязательно, чтобы сжимался каждый буфер. Условия для выбора числа кортежей в буфере (58) или (59) важны потому, что эффект от сжатия части буферов, число кортежей в которых удовлетворяет (58) или (59), должен превышать эффект от растяжения отдельных буферов, причем разбиение файла на буферы при заданных I и P дожно обеспечивать существование решения уравнения (27). В этом случае, когда не все буферы сжимаются, а только часть из них, нетрудно сформулировать условие сжатия в виде

$$LI\log_2 P + \sum_{l=1}^{L} m_l \log_2\left(\frac{m_l}{e}\right) < n\sum_{l=1}^{L} m_l \tag{60}$$

для α из (32), или в виде

$$L(2^P/L + I\log_2 P) + \sum_{l=1}^{L} m_l \log_2\left(\frac{m_l}{e}\right) < n\sum_{l=1}^{L} m_l \tag{61}$$

для α из (33).

3. Код поля кратности и поля порядка

Говоря о сжатии файла за счет выполнения условий (20) или (21), нужно помнить, что обеспечение этих условий — вспомогательный инструмент для увеличения коэффициента сжатия, а главным адаптирующим фактором является такое разбиение файла на буферы с длинами m_l, $l = 1, 2, \ldots, L$, чтобы при минимальных P и I уравнение (27) имело решения. Поэтому задача об обеспечении выполнения условий (20) или (21) должна ставиться следующим образом: *для данного разбиения файла подобрать адаптирующие факторы так, чтобы выполнялись условия* (20) *или* (21)*, но без изменения* m_l*, т. к. это изменение может нарушать существование решений уравнения* (27) *или, сохраняя это существование, увеличивать* P *или* I.

Алгоритм построения кода поля кратности описан в двух вариантах в работах [5, 6]. Условие (20) является приближенным. Более точное условие согласно [6] имеет вид

$$L_2 < \log_2 D^s_{m-s} + \log_2 \frac{p!}{\prod\limits_{j=1}^{p} l_j!}\ , \tag{62}$$

где p — число групп чисел повторов, содержащих одинаковые числа повторов, а l_j — число одинаковых чисел повторов стандартной формы, причем $j = 1, 2, \ldots, p$ и выполнено условие

$$s = \sum_{j=1}^{p} l_j. \tag{63}$$

Для оценки второго члена в правой части (62) можно воспользоваться формулой Стирлинга, считая, что $l_1 = l_2 = \ldots = l_s$ и $p = s$. Тогда будем иметь

$$\log_2 \frac{p!}{\prod\limits_{j=1}^{p} l_j!} = \log_2 s! \approx s(\log_2 s - \log_2 e). \tag{64}$$

Для оценки первого члена в правой части (62) можно воспользоваться асимптотическими формулами, полученными в [3]:

$$\log_2 D^s_{m-s} \approx 2.82(sm^{1/2})^{1/2} \text{ при } m \gg s \gg 1, \tag{65}$$

$$\log_2 D^s_{m-s} \approx 2.56m^{1/2} \text{ при } m \sim s \gg 1. \tag{66}$$

Из (62), (64), (65) получим, что

$$L_2 < 2.82\ (sm^{1/2})^{1/2} + s(\log_2 s - \log_2 e), \tag{67}$$

а из (62), (64), (66) —

$$L_2 < 2.56\ m^{1/2} + s(\log_2 s - \log_2 e). \tag{68}$$

Единственным фактором адаптации для кода поля кратности, является способ лексикографичекого упорядочения таблицы стандартных форм [6]. Этот способ может быть изменен, если задать перестановку из s чисел кратности $n_1, n_2, \ldots, n_s$, а именно:

$$i_1, i_2, \ldots, i_s. \tag{69}$$

При этом номер в этой таблице N_1 и номер в таблице перестановок N_2 изменятся так: $N_1 \to \tilde{N}_1$ и $N_2 \to \tilde{N}_2$. Длина кода поля кратности для каждого буфера есть $\log_2(N_1 N_2)$. Тогда если для всех буферов выполнено условие

$$\sum_{l=1}^{L} \log_2(\tilde{N}_1^l \tilde{N}_2^l) < 2.82\ (sm^{1/2})^{1/2} L \tag{70}$$

в случае (65) и

$$\sum_{l=1}^{L} \log_2(\tilde{N}_1^l \tilde{N}_2^l) < 2.56\ m^{1/2} \tag{71}$$

в случае (66), так как в код поля кратности нужно включить перестановку (69) и это потребует $\log_2 s(\log_2 s - \log_2 e)$ битов, то и условия (67) или (68) будут выполнены, и код поля кратности будет меньше.

Что касается условия (21) сокращения длины кода поля порядка, то поскольку этот код тоже, аналогично коду поля кратности, задается номером в лексикографичекски упорядоченной таблице перестановок [7], то для него тоже существует фактор адаптации, связанный со способом упорядочения этой таблицы, который может быть задан соответствующей перестановкой

$$j_1, j_2, \ldots, j_s. \tag{72}$$

Однако, для этого поля есть еще одна возможность адаптации кода к файлу. Действительно, порядок, в котором появляются кортежи при декодировании поля принадлежности, в определенной степни зависит от способа решения булева уравнения из поля принадлежности. Так как булевы уравнения легко решать последовательно придавая n булевым переменным $x_1, x_2, \ldots, x_n$ значения равные 0 или 1, то существует ξ, равное

$$\xi = 2^n n! \tag{73}$$

разных способов такого решения, поскольку можно начинать решения с любой из n булевых переменных и придать ей сначала значение 0, потом

1, или наоборот. Задание способа решения уравнения потребует дополнительно к коду порядка ΔL_3 битов:

$$\Delta L_3 = n(\log_2 n + 1 - \log_2 e). \tag{74}$$

При этом номер N в таблице перестановок поля кратности изменится: $N \to \tilde{N}$. Если будет выполнено условие

$$\log_2 \tilde{N} + \Delta L_3 < \log_2 N \tag{75}$$

или, согласно (73),

$$\log_2 \tilde{N} + n(\log_2 n + 1 - \log_2 e) < \log_2 N \tag{76}$$

для каждого буфера, то условие (20) будет выполняться, и коэффициент сжатия станет увеличиваться. Условие (75) можно переписать и для всего файла

$$\sum_{l=1}^{L} \log_2 \tilde{N}_l + n(\log_2 n + 1 - \log_2 e) < \sum_{l=1}^{L} \log_2 N_l, \tag{77}$$

если способ решения булева уравнения из кода поля принадлежности может быть подобран одним и тем же способом для всех буферов сразу.

4. Факторы адаптации

Соберем вместе факторы адаптации кода к файлу, рассмотренные выше.

1. n — длина кортежа или число булевых переменных $x_1, x_2, \ldots, x_n$ от которых зависят булевы полиномы поля принадлежности. Выбор n определяется архитектурой компьютера и, чтобы избежать деления ячейки, n должно быть кратно 8: $n = 8, 16, 24, 32, \ldots$.

2. Число буферов в файле L и число кортежей m_l, $l = 1, 2, \ldots, L$, т. е. разбиение файла на кортежи, при заданном n. Этот фактор адаптации определяется требованием существования решения уравнения (27) и условиями (60) или (61).

3. Число порождающих булевых полиномов P и длина кортежа из этих порождающих I. В случае условия (61), $P = 2, 3, \ldots, n-1$, а I определяется (30).

4. Способ упорядочения таблицы стандартных форм чисел повторов поля кратности. Задается перестановкой (69) и должен удовлетворять условию (70) или (71).

5. Способ упорядочения таблицы перестановок поля порядка. Задается перестановкой (72) и должен удовлетворять условию (74), где

$$\Delta L_3 = s(\log_2 s - \log_2 e). \tag{78}$$

6. Способ решения булева уравнения. Задается номером в таблице перестановок на n булевых переменных и кортежем из n нулей и единиц,

показывающим, какое значение придать булевой переменной сначала, а какое — потом. Должен удовлетворять условию (76).

Построение алгоритмов для оптимального выбора факторов адаптации требует дальнейших исследований, причем главной задачей является построение алгоритма для выбора L и m_l.

Список использованной литературы

1. *Толстопятов А. А.* О структуре дискретной информации и общих условиях ее сжатия // Вестник ИвГУ. – 2002. – Вып. 3. – С. 80–82.
2. *Толстопятов А. А.* О возможности использования булевых уравнений для сжатия файлов // Вестник ИвГУ. – 2003. – Вып. 3. – С. 82–84.
3. *Толстопятов А. А.* Вычисление длины поля кратности при булевом сжатии файлов // Вестник ИвГУ. – 2004. – Вып. 3. – С. 71-76.
4. *Толстопятов А. А., Хашин С. И.* Алгоритм построения поля порядка при булевом сжатии // Вестник ИвГУ. – 2004. – Вып. 3. – С. 139–143.
5. *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.
6. *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.
7. *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.

Поступила в редакцию 8.12.2007.