Математика и её приложения: ЖИМО. 2008. Выпуск 1 (5). С. 89 – 96.

УДК 512.54

А. А. Толстопятов[1]

Сравнение булева сжатия со стандартным

Ключевые слова: булево сжатие, частотное сжатие.

Рассмотрены частотные методы сжатия файлов по сравнению с булевым сжатием. Сравнение проводится по параметрам разбиения файла и ожидаемым коэффициентам сжатия.

We consider frequency methods of file compression and compare theirs with Boolean compression method by parametres of splitting of the file and by expected aspect ratio.

Предложенные К. Шенноном [5] методы сжатия информации, основанные на различии частот появления кортежей, на которые разбит файл, были развиты в нескольких направлениях [1]. Это замена кода Шеннона – Фано на оптимальный код Хаффмана, арифметическое кодирование, позволяющее кодировать отдельный кортеж нецелым числом битов, метод скользящего словаря, позволяющий разбивать файл на кортежи неравной длины, причем максимальная длина кортежа может быть такова, что для кортежей такой длины невозможно набрать достоверной статистики, а значит стабилизировать частоты появления таких кортежей [1]. Но все эти методы основаны на существовании в файле повторяющихся кортежей и отличаются друг от друга только тем, что называется повтором и способом построения кода.

В [3] был предложен иной подход к сжатию информации, основанный на том, что кортежи объединяются в буферы, и кортежи, входящие в каждый буфер, рассматриваются как решения булева уравнения, причем в код буфера входят коэффициенты этого уравнения и коды полей кратности и порядка, задающие кратность вхождения каждого кортежа в буфер и расстановку кортежей с учетом их кратности внутри каждого буфера. В [2] было рассмотрено сравнение стандартного и булева кодирования. Но там был взят предельный случай, когда при булевом кодировании каждый буфер набирается до тех пор, пока в него не войдут все 2^n кортежей, где n — длина кортежа. Поскольку число кортежей, входящих в один буфер, ограничено сверху условием сжатия буфера, то, как показано в [2], для разных типов реальных файлов это число оказывается превышенным, а значит при этом очень специализированном и совсем не обязательном, хо-

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

тя и очень упрощающем кодирование условии вхождения всех возможных кортежей в каждый буфер, сжатия не происходит. В настоящей работе будет осуществелено сравнение по коэффициенту сжатия стандартного частотного сжатия и булева сжатия. Поскольку целью работы является выявление принципиальных моментов такого сравнения, то мы ограничимся простейшим случаем, когда файл разбивается на кортежи равной длины. Более общий случай, который требуется, например, в методе скользящего словаря, сильно усложняет такое сравнение, но по существу ничего не меняет.

1. Постановка задачи

Пусть N_Φ — длина файла, и файл разбит на кортежи длиной n. Тогда, обозначая через m число кортежей в файле, будем иметь:

$$m = \frac{N_\Phi}{n}. \tag{1}$$

Будем предполагать, что N_Φ настолько велико, что в файл входят все 2^n разных кортежей. При частотном сжатии коэффициент сжатия отдельного кортежа зависит только от одной величины — энтропии H. А значит и коэффициент сжатия всего файла зависит только от энтропии. Энтропия в свою очередь зависит только от вероятности p_i появления i-го кортежа в файле. Значит коэффициент сжатия всего файла определяется распределением вероятностей $p_i,\ i = 1, 2, \ldots, 2^n$.

При булевом сжатии m кортежей объединяются в L буферов. Пусть m_l — число кортежей, входящих в l-й буфер, а s_l — число различных кортежей входящих в l-й буфер, $l = 1, 2, \ldots, L$. Ясно, что

$$\sum_{l=1}^{L} m_l = m. \tag{2}$$

Если в l-м буфере i-й кортеж встречается n_1^l раз, второй — n_2^l раз, ..., s_l-ый — n_1^l раз, то коэффициент сжатия l-го буфера k_l будет зависеть не только от m_l и s_l, а еще и от чисел n_k^l, $l = 1, 2, \ldots, L,\ k = 1, 2, \ldots, s_l$, которые назовем числами повторов.

Задача, рассматриваемая в настоящей работе, заключается в том, чтобы сравнить коэффициенты сжатия файла при частотном и булевом сжатии и выяснить, при каких ограничениях на параметры разбиения файла какой из коэффициентов сжатия может быть больше. Заметим, что если при частотном сжатии при заданной длине файла N_Φ и фиксированном n вообще отсутствуют факторы адаптации, т. е. коэффициент сжатия определяется однозначно, то при булевом сжатии такой фактор адаптации

есть. Им являются числа m_l, $l = 1, 2, \ldots, L$, задающие способы объединения кортежей в буферы. От этих способов зависит коэффициент сжатия файла и, меняя эти числа, можно изменить коэффициент сжатия.

2. Коэффициент сжатия при частотном кодировании

Если через N_k обозначить длину кода, то коэффициент сжатия k есть:

$$k = \frac{N_{\Phi}}{N_k}. \tag{3}$$

При частотном кодировании файла с энтропией H, $N_k = mH$. Поэтому из (3) получим

$$k_{\text{ч}} = \frac{n}{H}. \tag{4}$$

Так как энтропия H есть

$$H = -\sum_{i=1}^{2^n} p_i \log_2 p_i, \tag{5}$$

то она удовлетворяет неравенству

$$0 < H \leq \log_2 2^n = n, \tag{6}$$

причем максимальное значение H достигает, если все вероятности p_i появления i-го кортежа в файле одинаковы и равны

$$p_i = 2^{-n}. \tag{7}$$

Поэтому $k \geq 1$.

Обозначая через n_i, $i = 1, \ldots, 2^n$ числа повторов i-ых кортежей в файле, можно выразить p_i через n_i:

$$p_i = \frac{n_i}{2^n}. \tag{8}$$

Используя (8), перепишем (5) в виде

$$H = \frac{1}{2^n}\left(n\sum_{i=1}^{2^n} n_i - \sum_{i=1}^{2^n} n_i \log_2 n_i\right). \tag{9}$$

Если учесть, что

$$\sum_{i=1}^{2^n} n_i = m, \tag{10}$$

то вместо (9) будем иметь:

$$H = \frac{1}{2^n}\left(mn - \sum_{i=1}^{2^n} n_i \log_2 n_i\right). \tag{11}$$

А тогда из (4) и (11) найдем, что

$$k_{\text{Ч}} = \frac{2^n \cdot n}{mn - \sum\limits_{i=1}^{2^n} n_i \log_2 n_i}. \tag{12}$$

Форма (12) представления k является удобной для сравнения с коэффициентом сжатия при булевом кодировании.

3. Коэффициент сжатия при булевом кодировании

При булевом кодировании m кортежей, на которые разбит файл, объединяются в L буферов, каждый из которых содержит по m_l кортежей, где $l = 1, \ldots, L$, и характеризуется своим коэффициентом сжатия k_l.

Если кодировать поле кратности не двумя, а одним числом, как предложено в [4], то длина кода l-го буфера N_k^l, считая, что на каждый буфер приходится равные части длины кода общего поля, будет равна

$$N_k^l = \frac{1}{L}(2^I + 2^n P) + I \log_2 P + \log_2 \frac{C_{m_l-1}^{s_l-1} m_l!}{\prod\limits_{k=1}^{s_l} n_k^l!}, \tag{13}$$

где в (13) первый член — часть длины кода общего поля, приходящаяся на l-ый буфер, второй член — длина кода поля принадлежности, а третий — сумма длин поля кратности и порядка, причем через I обозначено число переменных кодирующего полинома, а через P — число порождающих булевых полиномов [5].

Если, следуя [5], ввести обозначения

$$\alpha = 2^I + 2^n P + IL \log_2 P, \tag{14}$$

$$\beta_l = \log_2 \frac{C_{m_l-1}^{s_l-1} m_l!}{\prod\limits_{k=1}^{s_l} n_k^l!}, \tag{15}$$

то (13) можно переписать так:

$$N_k^l = \frac{\alpha}{L} + \beta_l. \tag{16}$$

Тогда, обозначая через $N_{\text{б}}^l$ длину l-го буфера, получим, что

$$k_l = \frac{N_{\text{б}}^l}{N_k^l} = \frac{m_l n L}{\alpha + L\beta_l}. \tag{17}$$

Поскольку при частотном сжатии число $k_{\text{ч}}$ из (12) есть коэффициент сжатия всего файла, а при булевом сжатии кортежи объединяются в буферы, содержащие m_l кортежей, и каждый буфер характеризуется своим коэффициентом сжатия k_l из (17), то необходимо выразить коэффициент сжатия всего файла $k_{\text{б}}$ через k_l. Как показано в [5], это можно сделать двумя способами. Если ввести обозначение

$$\gamma_l = \frac{\beta_l}{\alpha}, \tag{18}$$

то справедлива формула

$$k_{\text{б}} = \frac{< k_L > + \sum\limits_{l=1}^{L} \gamma_l k_l}{1 + \sum\limits_{l=1}^{L} \gamma_l}, \tag{19}$$

где через $< k_L >$ обозначено среднеарифметическое коэффициентов сжатия отдельных буферов

$$< k_L >= \frac{1}{L} \sum_{l=1}^{L} k_l. \tag{20}$$

Второй способ требует введения величин

$$< W_l >= \frac{1 + L\gamma_l}{L\left(1 + \sum\limits_{l=1}^{L} \gamma_l\right)}, \tag{21}$$

причем W_l удовлетворяют условиям:

$$0 < W_l \leq 1, \tag{22}$$

$$\sum_{l=1}^{L} W_l = 1. \tag{23}$$

Тогда справедлива формула:

$$k_{\text{б}} = \sum_{l=1}^{L} W_l k_l. \tag{24}$$

4. Сравнение коэффициентов сжатия при частотном и булевом кодировании

Чтобы сравнить коэффициенты сжатия при частотном и булевом кодировании, нужно найти связь между числами повторов n_i, $i = 1, \ldots, 2^n$, во всем файле и числами повторов n_k^l, $k = 1, \ldots, s_l$, в l-ом буфере, $l = 1, \ldots, L$. Поскольку в l-ый буфер могут входить не все 2^n кортежей, а входящие s_l кортежей с числами повторов n_k^l, $k = 1, \ldots, s_l$, нумеруются числами от 1 до s_l, то изменим нумерацию этих кортежей следующим способом. Будем считать, что индекс i, нумерующий кортежи в l-ом буфере, пробегает значения от 1 до 2^n, а если j-ый кортеж в l-ый буфер не входит, то ему приписывается число повторов $n_j^l = 0$. Тогда

$$n_i = \sum_{l=1}^{L} n_i^l, \tag{25}$$

и формула (12) для $k_{\text{ч}}$ примет вид:

$$k_{\text{ч}} = \frac{2^n n}{mn - \sum\limits_{i=1}^{2^n} \left(\sum\limits_{l=1}^{L} n_i^l \log_2 \sum\limits_{l=1}^{L} n_i^l \right)}. \tag{26}$$

Записав k_l из (17) в виде

$$k_l = \frac{m_l n L}{\alpha + L\gamma_l} \tag{27}$$

и подставив (2) и (27) в (24), с учетом (2) и (18), получим следующее выражение для $k_{\text{б}}$:

$$k_{\text{б}} = \frac{mn}{\alpha + \sum\limits_{l=1}^{L} \beta_l}. \tag{28}$$

Поскольку нас интересует случай, когда булево кодирование эффективнее частотного, т. е. когда выполняется неравенство

$$k_{\text{б}} > k_{\text{ч}}, \tag{29}$$

то, подставляя (26) и (28) в (29), получим

$$\sum_{l=1}^{L} \beta_l < \frac{m}{2^n} \left[mn - \sum_{i=1}^{2^n} \left(\sum_{l=1}^{L} n_i^l \log_2 \sum_{i=1}^{L} n_i^l \right) - \alpha \right]. \tag{30}$$

Если, как оговорено выше, приписывать кортежам, не входящим в буфер нулевые числа повторов, то формула (15) для β_l может быть записана так:

$$\beta_l = \log_2 \frac{C_{m_l-1}^{s_l-1} m_l!}{\prod\limits_{k=1}^{2^n} n_k^l!}. \tag{31}$$

Тогда, подставляя (31) в (30), получим

$$\sum_{i=1}^{2^n} \sum_{l=1}^{L} \log_2 \frac{n_i^l!}{\left(\sum\limits_{k=1}^{L} n_i^k\right)^{n_i^l}} > \log_2 \prod_{k=1}^{s_l} \left(C_{m_l-1}^{s_l-1} m_l!\right) - \frac{(mn-\alpha)m}{2^n}. \tag{32}$$

Записав правую часть (32) в виде суммы

$$\log_2 \prod_{k=1}^{s_l} \left(C_{m_l-1}^{s_l-1} m_l!\right) - \frac{(mn-\alpha)m}{2^n} = \sum_{l=1}^{L} \left(\log_2 \left(C_{m_l-1}^{s_l-1} m_l!\right) - \frac{(mn-\alpha)m_l}{2^n}\right), \tag{33}$$

меняя порядок суммирования в левой части (32) и подставляя (33) в (32), получим

$$\sum_{l=1}^{L} \left\{ \sum_{i=1}^{2^n} \log_2 \frac{n_i^l!}{\left(\sum\limits_{k=1}^{L} n_i^k\right)^{n_i^l}} + \frac{(mn-\alpha)m_l}{2^n} - \log_2 \left(C_{m_l-1}^{s_l-1} m_l!\right) \right\} > 0. \tag{34}$$

Заметим, что не все члены во внешней сумме (34) могут быть больше нуля. Поэтому условие (34) может выполняться и тогда, когда суммируются как положительные так и отрицательные слагаемые в фигурных скобках в (34), а поэтому внешнюю сумму нельзя снять, потребовав выполнения неравенства (34) для отдельных слагаемых.

Неравенство (34) содержит только величины, характеризующие разбиение файла на кортежи и объединение этих кортежей в буферы при булевом кодировании. Поэтому для любого разбиения файла оно может быть проверено. Если (34) выполняется, то выполняется (29), а значит булево сжатие эффективнее частотного. Если (34) не выполняется, то выполняется неравенство, противоположное (29), а значит эффективнее частотное сжатие.

Список использованной литературы

1. *Ватолин Д., Ратушняк А., Смирнов М., Юдин В.* Методы сжатия данных. Устройство архиваторов, сжатие изображений и видео. – М.: Диалог - МИФИ, 2002. – 384 с.

2. *Толстопятов А. А.* О структуре дискретной информации и общих условиях ее сжатия // Вестник ИвГУ. – 2002. – Вып. 3. – С. 80–82.

3. *Толстопятов А. А.* О возможности использования булевых уравнений для сжатия файлов // Вестник ИвГУ. – 2003. – Вып. 3. – С. 82–84.

4. *Толстопятов А. А.* Возможность кодирования поля кратности одним числом // Математика и ее приложения: Журн. Иванов. матем. об-ва. – 2008. – Вып. 1(5). – С. 43–52.

5. *Толстопятов А. А.* Алгоритм разбиения файла на буферы при булевом сжатии // Математика и ее приложения: Журн. Иванов. матем. об-ва. – 2008. – Вып. 1(5). – С. 77–88.

6. *Шеннон К.* Математическая теория связи // Работы по теории информации и кибернетике. – М.: ИЛ, 1963. – С. 243–332.

Поступила в редакцию 29.12.2008.