УДК 621.391

Н. Н. ПОНОМАРЕНКО, Н. В. КОЖЕМЯКИНА

Национальный аэрокосмический университет им. Н. Е. Жуковского «ХАИ», Украина

РЕКУРСИВНОЕ ГРУППОВОЕ КОДИРОВАНИЕ С КОЛИЧЕСТВОМ И РАЗМЕРАМИ ГРУПП, НЕ ЗАВИСЯЩИМИ ОТ КОДИРУЕМЫХ ДАННЫХ

Рассмотрена задача рекурсивного группового кодирования данных с целью устранения их статистической избыточности. Показано, что в ряде случаев, например, при динамическом варианте кодирования, целесообразно использовать постоянные размеры и количество групп вместо того, чтобы вычислять их адаптивно по отношению к кодируемому тексту. Предложено несколько вариантов сочетаний размеров групп, проведен сравнительный анализ эффективности их использования при кодировании. Предложен адаптивный метод определения количества итераций кодирования. Показано, что для стандартных тестовых наборов данных предлагаемая модификация обеспечивает коэффициенты сжатия, сравнимые с обычным рекурсивным групповым кодированием, а при сжатии небольших файлов может даже превосходить его.

Ключевые слова: *сжатие данных, энтропийное кодирование, рекурсивное групповое кодирование.*

Введение

Задачи сжатия данных сохраняют высокую актуальность последние несколько десятков лет [1, 2]. При этом внимание уделяется как разработке высокоуровневых методов сжатия, таких как предсказание по частичному совпадению [3] или сжатие изображений и видео [2, 4, 5], так и элементарных методов, предназначенных для устранения статистической избыточности, таких как арифметическое кодирование [6] или кодирование Хаффмана [7].

Одной из наиболее перспективных альтернатив арифметическому кодированию является рекурсивное групповое кодирование (РГК) [8, 9]. Главным достоинством РГК является его способность эффективно кодировать символы сверхбольших алфавитов [8], что очень актуально при кодировании мультимедийных данных, например, квантованных коэффициентов дискретного косинусного преобразования при сжатии изображений и видео [5, 10, 11]. Кроме того, РГК является вычислительно очень простым, так как в основном цикле кодирования использует только операции логического "или" и сдвига. РГК обладает одной из самых быстрых скоростей кодирования данных, что при степенях сжатия сравнимых, а часто и существенно более высоких, чем для арифметического кодирования, делает его привлекательным для кодирования больших потоков однородных данных, например, для сжатия трафика в сетях передачи данных.

Недостатком РГК является необходимость предварительно оценить частоты встречаемости символов алфавита в тексте, затем сформировать супербуквы (группы символов, близкие по частотам встречаемости) и сохранить их состав в сжатых данных. Это сужает сферу применения РГК кодированием статистически однородных данных.

Целью данной статьи является разработка модификации РГК с универсальными сочетаниями размеров супербукв, которые могли бы эффективно использоваться при кодировании любых текстов. Это позволило бы использовать РГК вместо арифметического кодирования в составе высокоуровневых методов сжатия, в которых заранее невозможно оценить частоты встречаемости символов текста.

Кроме того, предлагается критерий остановки итераций кодирования, который в отличие от стандартного РГК, в котором кодирование останавливается при длине текста меньшей 200 [9], позволяет остановить кодирование, если дальнейшие итерации приведут к увеличению размера сжатых данных.

1. Метод использования постоянного сочетания размеров групп

В РГК на каждой итерации кодирования для каждого k-го символа алфавита (обозначим общее число символов в алфавите как K) оценивается частота его встречаемости в кодируемом тексте p_k. Затем все символы алфавита разделяются на S супербукв (групп символов) с близкими значениями p_k. После этого суффиксы символов (порядковые номера внутри супербукв) сохраняются в сжатом файле, а префиксы (номера супербукв) попарно объединяются и, сформированный из них новый текст (в два раза короче исходного) подается на следующую итерацию кодирования. Чтобы размерность алфавита от итерации к итерации не возрастала должно выполняться условие $K \geq S^2$.

Увеличение длины кода из-за объединения

© Н. Н. Пономаренко, Н. В. Кожемякина

символов в супербукву в относительных единицах выражается как

$$\Delta = -p_E(-\log_2 p_E + \log_2 M)/\sum_{k=1}^{M}(p_k \log_2 p_k), \quad (1)$$

где M - число символов, объединенных в супербукву, p_E - суммарная вероятность этих символов $\sum_{k=1}^{M} p_k$.

В [9] предложен адаптивный алгоритм разбиения алфавита на супербуквы для заданного текста, обеспечивающий Δ не выше заданного порога, для работы которого нужно знать p_k для всех символов.

Если p_k неизвестны и нужно задать массив размеров супербукв $L = \{L_1, L_2, ..., L_S\}$, то для минимизации Δ необходимо выполнение условия

$$L_j \le L_{j+1/} \quad (2)$$

Тогда символы алфавита, упорядоченные по убыванию p_k, можно распределить между супербуквами следующим образом. Первые L_1 символов алфавита с самыми большими p_k будут соответствовать первой супербукве. Следующие L_2 символов алфавита будут соответствовать второй супербукве и т.д. Выполнение условия $L_i \le L_{i+1}$ позволит обеспечить меньший размер супербукв для символов с большими p_k, то есть кодирование этих символов будет осуществляться со статистической избыточностью меньшей, чем для символов с меньшими p_k.

Структурная схема кодирования для модифицированного таким образом метода РГК приведена на рис. 1.

Для $K=16^2$ можно рассмотреть, например, такие варианты массива L:

L1 = {1, 1, 2, 4, 8, 16, 32, 64, 128},
L2 = {1, 1, 1, 1, 2, 2, 4, 4, 8, 8, 16, 16, 32, 32, 64, 64},
L3 = {1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 241},
L4 = {1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 4, 16, 32, 64, 128}.

Для варианта L1 количество супербукв равно восьми, для вариантов L2, L3 и L4 количество супербукв равно 16. Все варианты учитывают требование (2), причем символ с самым большим значением p_k во всех вариантах предлагается выделить в отдельную группу, а для последних групп, состоящих из редко встречающихся в тексте символов, предлагается задавать большие размеры (64, 128, 241). Избыточность кода для редко встречающихся символов будет большой, но это не окажет большого влияния на общий объем сжатого текста.

При использовании постоянного S=16, количество символов в тексте на второй и последующей итерациях будет близким к 256.

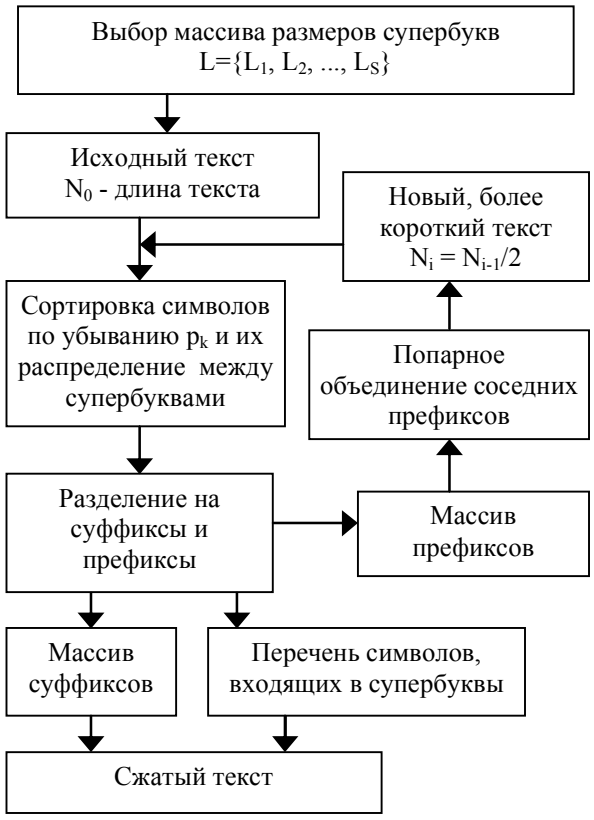

Рис. 1. Структурная схема кодирования для РГК с постоянными размерами супербукв

Таким образом, на каждой итерации для сохранения перечня символов, входящих в супербуквы, потребуется около 256 байт. При этом правило остановки итераций, если $N_i<200$ (использующееся для стандартного РГК), в данном случае может привести к сильной избыточности кода для нескольких последних итераций. Предложим простое правило остановки итераций, которое будет лишено этого недостатка. Пусть результаты i-й итерации сжатия сохраняются в сжатом файле только, если выполняется условие:

$$N_i T \ge F_i + 1 + K_i + N_i/2, \quad (3)$$

где N_i – длина исходного текста для i-ой итерации, F_i – объем памяти, занимаемой массивом суффиксов для i-ой итерации, $N_i/2$ – объем памяти, занимаемый попарно объединенными номерами супербукв, 1 – один байт, необходимый для хранения числа разных символов, K_i – объем памяти, необходимый для хранения перечня символов, входящих в супербуквы, T – коэффициент, больший или равный единице, позволяющий допускать некоторое превышение объема сжатого текста над объемом исходного текста. В данной работе предлагается задаватьT = 1,2.

Если условие (3) не выполняется, то исходный текст для данной итерации должен быть сохранен в сжатых данных в неизмененном виде.

3. Анализ эффективности предложенной модификации РГК

Сравним эффективность кодирования для предлагаемой модификации РГК с постоянными размерами супербукв для вариантов L1, L2, L3, L4. Кроме того, рассмотрим все эти варианты как со стандартным правилом остановки итераций (СПОИ), так и с предложенным (ППОИ). Для сравнения приведем результаты сжатия для стандартного РГК с адаптивным выбором состава супербукв.

Для сравнительного анализа будем использовать тестовые наборы Calgary corpus test files [12], Canterbury corpus test files [13], PICST [14], а также использованные в [9] наборы случайных чисел с нормальным законом распределения и квантованных коэффициентов ДКП блоков изображений. В таблицах приведено число бит в сжатом тексте на каждый символ исходного текста.

Таблица 1

Тестовый набор Canterbury corpus test files

Файл	РГК	СПОИ				ППОИ			
		L1	L2	L3	L4	L1	L2	L3	L4
alice29.txt	4,60	4,62	4,45	5,09	4,42	4,62	4,44	5,09	4,41
asyoulik.txt	4,82	4,95	4,77	5,38	4,73	4,95	4,75	5,38	4,71
cp.html	5,44	5,47	5,42	6,56	5,39	5,47	5,32	6,56	5,36
fields.c	5,33	5,32	5,36	5,93	5,32	5,32	5,28	5,85	5,24
grammar.lsp	5,48	5,31	5,5	5,59	5,32	5,31	5,29	5,59	5,18
kennedy.xls	3,15	3,21	2,67	2,30	2,41	3,21	2,66	2,30	2,41
lcet10.txt	4,60	4,72	4,48	5,28	4,47	4,72	4,48	5,28	4,46
plrabn12.txt	4,47	4,60	4,37	5,19	4,35	4,6	4,37	5,19	4,34
ptt5	0,97	0,98	0,94	1,17	0,95	0,98	0,94	1,17	0,95
sum	5,23	5,30	5,25	6,00	5,24	5,30	5,20	5,98	5,22
xargs.1	5,67	5,67	6,01	6,36	5,92	5,67	5,54	6,36	5,49

Для данных табл. 1 и 2 для всех файлов, кроме OBJ1, предложенная модификация (варианты L2 и L4) с ППОИ обеспечивает лучшие результаты, чем РГК от 1% до 20% (файл kennedy.xls).

Таблица 2

Тестовый набор Calgary corpus test files

Файл	РГК	СПОИ				ППОИ			
		L1	L2	L3	L4	L1	L2	L3	L4
BIB	5,19	5,18	4,91	5,81	4,86	5,18	4,9	5,81	4,83
BOOK1	4,47	4,61	4,41	5,28	4,37	4,61	4,41	5,28	4,37
BOOK2	4,76	4,89	4,69	5,49	4,67	4,89	4,69	5,49	4,66
GEO	4,71	4,83	4,57	4,74	4,52	4,83	4,56	4,74	4,50
NEWS	5,19	5,29	5,13	5,89	5,12	5,29	5,13	5,89	5,12
OBJ1	5,88	5,87	6,00	6,44	6,00	5,87	5,97	6,44	5,97
OBJ2	6,02	6,08	5,72	6,42	5,65	6,08	5,71	6,42	5,64
PAPER1	5,06	5,12	5,09	5,91	5,08	5,12	5,06	5,90	5,04
PAPER2	4,65	4,72	4,60	5,46	4,59	4,72	4,57	5,46	4,56
PIC	0,96	0,98	0,94	1,17	0,95	0,98	0,94	1,17	0,95
PROGC	5,33	5,37	5,35	6,18	5,32	5,37	5,30	6,18	5,29
PROGL	4,82	4,83	4,58	5,54	4,49	4,83	4,56	5,53	4,47
PROGP	4,88	4,92	4,79	5,64	4,78	4,92	4,77	5,64	4,76
TRANS	5,56	5,59	5,32	6,40	5,30	5,59	5,30	6,40	5,28

Таблица 3

Результаты сжатия для тестового набора PICST

Файл	РГК	СПОИ				ППОИ			
		L1	L2	L3	L4	L1	L2	L3	L4
d2.dat	7,87	8,01	7,90	8,35	7,97	8,00	7,90	8,00	8,00
d4.dat	7,66	7,85	7,74	8,36	7,79	7,85	7,74	8,00	7,78
d8.dat	7,37	7,62	7,52	8,38	7,54	7,62	7,51	8,00	7,54
d16.dat	7,00	7,32	7,19	8,37	7,24	7,32	7,19	8,00	7,23
d32.dat	6,65	6,91	6,79	8,31	6,84	6,91	6,79	8,00	6,84
d64.dat	6,14	6,48	6,35	8,12	6,41	6,48	6,35	8,12	6,41

Таблица 4

Гауссов шум

Файл	РГК	СПОИ				ППОИ			
		L1	L2	L3	L4	L1	L2	L3	L4
odn25	4,45	4,57	4,49	4,92	4,64	4,57	4,48	4,67	4,62
odn100	5,45	5,59	5,49	6,86	5,59	5,59	5,48	6,86	5,59
odn400	6,45	6,57	6,48	8,3	6,55	6,57	6,47	8,00	6,54

Таблица 5

Квантованные коэффициенты ДКП изображений

Файл	РГК	СПОИ				ППОИ			
		L1	L2	L3	L4	L1	L2	L3	L4
barb10	1,68	1,77	1,70	1,96	1,72	1,77	1,69	1,96	1,71
barb50	0,56	0,58	0,56	0,65	0,56	0,58	0,56	0,65	0,56
len10	1,29	1,40	1,35	1,56	1,36	1,40	1,34	1,56	1,35
len50	0,34	0,36	0,34	0,40	0,34	0,36	0,34	0,40	0,34

Для набора PICST, Гауссова шума и коэффициентов ДКП степень сжатия для ППОИ (варианты L2 и L4) хуже, чем для РГК на 1-5%.

Заключение

Предложенная в работе модификация РГК с фиксированными размерами групп и правилом остановки итераций обеспечивает сравнимую с РГК степень сжатия. При этом на небольших файлах может достигаться выигрыш до 20%, в то время как в статистически однородных файлах с алфавитами большой размерности модификация проигрывает стандартному РГК в степени сжатия всего 1-5%.

Литература

1. Salomon, D. Data Compression - The Complete Reference [Text] / D. Salomon. – Springer-Verlag. – 2004. – 898 p.

2. Методы сжатия данных [Текст] : учебно-справочное издание / Д. Ватолин, А. Ратушняк, М. Смирнов, В. Юкин. – М. : Диалог-Мифи, 2002. – 383 с.

3. Cleary, J. Data compression using adaptive coding and partial string matching [Text] / J. Cleary, I. Witten // IEEE Transactions on Communications. – April, 1984. – Vol. COM-32. – P. 396-402.

4. ADCTC: A new high quality DCT based coder for lossy image compression [Electronic resource] / N.

Ponomarenko, V. Lukin, K. Egiazarian, J. Astola. – 80 Min / 700 MB. CD ROM Proceedings of LNLA. – Switzerland, August – 2008. – 6 p. – 1 electronic optical disc (CD-ROM).

5. Bazhyna, A. V. Efficient bit-planes based method for compression of 3D-DCT coefficients [Text] / A. V. Bazhyna, K. O. Egiazarian, N. N. Ponomarenko // Proceedings of Picture Coding Symposium, Lisboa, Portugal, 7-9 November, 2007. –4 p.

6. Rissanen, J. Generalized kraft inequality and arithmetic coding [Text] / J. Rissanen // IBM J. Res. Develop. – May, 1976. – Vol. 20. – P. 198-203.

7. Huffman, D. A. A method for the construction of minimum-redundancy codes [Text] / D. A. Huffman // Proc. Inst. Radio Eng. – September, 1952. – Vol. 40, N 9. – P. 1098-1101.

8. Fast recursive coding based on grouping of Symbols [Text] / N. Ponomarenko, V. Lukin, K. Egiazarian, J. Astola // Telecommunications and Radio Engineering. – 2009. – Vol. 68, N 20. – P. 1857-1863.

9. Lossy compression of images without visible distortions and its application [Text] / V. Lukin, M. Zriakhov, N. Ponomarenko S. Krivenko, M. Zhenjiang // IEEE 10th International Conference on Signal Processing (ICSP). – 2010. – P. 698-701.

10. Пономаренко, Н. Н. Метод энтропийного рекурсивного группового кодирования [Текст] / Н. Н. Пономаренко, Н. В. Кожемякина, В. В. Лукин // Радіоелектронні і комп'ютерні системи. – 2014. – № 3 (67). – С. 20-26.

11. Quasi-optimal compression of noisy optical and radar images [Text] / V. Lukin, N. Ponomarenko, M. Zriakhov, A. Zelensky, K. Egiazarian // Image and Signal Processing for Remote Sensing XII, Stockholm, Sweden. – 2006. – Vol. 6365. – P. 1-11.

12. Calgary corpus test files [Electronic resource] : Archive Comparison Test, website. – Access mode: http://compression.ca/act/act-files.html. – Access date 25.05.2015. – Title by screen.

13. Canterbury corpus test files [Electronic resource] : Archive Comparison Test, website. – Access mode: http://corpus.canterbury.ac.nz. – Access date 25.05.2015. – Title by screen.

14. Means and results of efficiency analysis for data compression methods applied to typical multimedia data [Text] / N. Kozhemiakina, N. Ponomarenko, V. Lukin, K. Egiazarian, J. AStola // IEEE First International Scientific-Practical Conference Problems. – 2014. – P. 12-14.

Поступила в редакцию 12.05.2015, рассмотрена на редколлегии 18.06.2015

РЕКУРСИВНЕ ГРУПОВЕ КОДУВАННЯ З КІЛЬКІСТЮ І РОЗМІРАМИ ГРУП, ЩО НЕ ЗАЛЕЖАТЬ ВІД ДАНИХ, ЩО КОДУЮТЬСЯ

М. М. Пономаренко, Н. В. Кожемякіна

Розглянуто задачу рекурсивного групового кодування даних з метою усунення їх статистичної надмірності. Показано, що в ряді випадків, наприклад, при динамічному варіанті кодування, доцільно використовувати постійні розміри і кількість груп замість того, щоб обчислювати їх адаптивно по відношенню до тексту, що кодується. Запропоновано декілька варіантів поєднань розмірів груп, проведено порівняльний аналіз ефективності їх використання при кодуванні. Запропоновано адаптивний метод визначення кількості ітерацій кодування. Показано, що для стандартних тестових наборів даних запропонована модифікація забезпечує коефіцієнти стиснення, які порівняно зі звичайним рекурсивним груповим кодуванням, а при стисненні невеликих файлів може навіть перевершувати його.

Ключові слова: стиснення даних, ентропійне кодування, рекурсивне групове кодування.

RECURSIVE GROUP CODING WITH FIXED NUMBER AND SIZES OF GROUPS

N. N. Ponomarenko, N. V. Kozhemiakina

The task of elimination of data statistical redundancy by recursive group coding is considered. It is shown that in some cases, for example for dynamic coding, it is reasonable to use fixed sizes and number of groups instead of calculating them in adaptive manner with taking into account the encoded text. Several variants of combinations of group's sizes are proposed and comparative analysis of the effectiveness of their use in coding is carried out. The adaptive method for determining the number of iterations of coding is described. It is shown that for the standard test data sets the proposed modification provides compression ratios comparable to conventional recursive group coding while for small files even better compression ratios are reached.

Key words*:* data compression, entropy coding, recursive group coding.

Пономаренко Николай Николаевич - д-р техн. наук, доцент, профессор каф. приема, передачи и обработки сигналов, Национальный аэрокосмический университет им. Н. Е. Жуковского «Харьковский авиационный институт», Харьков, Украина, e-mail: nikolay@ponomarenko.info.

Кожемякина Надежда Владимировна – аспирант каф. приема, передачи и обработки сигналов, Национальный аэрокосмический университет им. Н. Е. Жуковского «Харьковский авиационный институт», Харьков, Украина, e-mail: nadejda_kozickaya@mail.ru.