УДК 621.391

Н. В. КОЖЕМЯКИНА, Н. Н. ПОНОМАРЕНКО

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

РЕКУРСИВНОЕ ГРУППОВОЕ КОДИРОВАНИЕ С ДИНАМИЧЕСКИМ ЧАСТОТНЫМ МОДЕЛИРОВАНИЕМ

Рассмотрена задача энтропийного кодирования данных с целью устранения в них статистической избыточности на основе рекурсивного группового кодирования. Рекурсивное групповое кодирование является более быстрой и в ряде случаев более эффективной альтернативой таким методам, как арифметическое кодирование и кодирование Хаффмана, за счет способности эффективно кодировать символы сверхбольших алфавитов. Однако его применение до сих пор ограничивалось статичным частотным моделированием, при котором для всего текста вычисляются частоты символов, а информация о размерах и составе групп сохраняется в файл вместе со сжатым текстом. В данной работе впервые предлагается модификация рекурсивного группового кодирования, позволяющая эффективно сжимать статистически неоднородные по частотным характеристикам тексты. Рассмотрена модель формирования тестовых наборов данных, предложена метрика для оценивания статистической неоднородности данных. Показано, что для рассмотренных тестовых данных предложенная модификация обеспечивает до 65% меньший объем сжатых данных, чем стандартный вариант рекурсивного группового кодирования.

Ключевые слова: *рекурсивное групповое кодирование, энтропийное кодирование, арифметическое кодирование, кодирование Хаффмана.*

Введение

Повсеместное использование телекоммуникационных технологий привело к росту числа передаваемых по сетям данных, прежде всего мультимедийных. При этом увеличение пропускной способности каналов передачи данных не успевает за ростом объема передаваемых данных, что делает актуальной разработку новых более эффективных методов сжатия этих данных [1]. При этом, наряду с разработкой высокоуровневых методов сжатия, таких как PPM [2] или кодирования Бэрроуза-Уилера [3], по-прежнему актуальной является и разработка "элементарных" методов устранения статистической избыточности в данных, таких как арифметическое кодирование (АК) [4] или кодирование Хаффмана (КХ) [5]. Такие методы используются в составе более сложных методов сжатия, обычно на последних этапах кодирования, когда необходимо устранить статистическую избыточность в данных.

Энтропийное рекурсивное групповое кодирование (ЭРГК) является относительно новым методом кодирования [6, 7] и относится к той же группе методов, что и АК с КХ. При этом ЭРГК обеспечивает более быстрое кодирование и декодирование данных при большей вычислительной простоте, чем у АК и КХ. Однако главным достоинством ЭРГК является способность более эффективно, чем АК, сжимать тексты со сверхбольшими алфавитами [8] при приблизительно такой же, как и у АК эффективности сжатия текстов однобайтных алфавитов. Это делает ЭРГК ценным инструментом в составе высокоуровневых методов сжатия мультимедийной информации, для которых характерно наличие символов больших алфавитов. Так, например, показано, что использование ЭРГК в составе стандарта JPEG вместо целой цепочки методов, включающей Zig-Zag сканирование, RLE-кодирование и КХ, позволяет увеличить степень сжатия на 5-10% [9]. Также хорошо ЭРГК зарекомендовало себя в сочетании с преобразованием Бэрроуза-Уилера и кодированием расстояний [10].

Недостатком ЭРГК, ограничивающим его применение только сжатием статистически частотно однородных данных, является наличие лишь варианта кодирования со статическим моделированием. При кодировании какого-либо текста для ЭРКГ необходимо разбить все символы по группам (супербуквам) в соответствии с частотой их встречаемости в тексте, а также сохранить количество, размеры и состав групп вместе со сжатыми данными. При динамическом частотном моделировании, когда таблица частот обновляется после кодирования каждого символа, происходит изменение количества групп и их состава, что для ЭРГК приводит к резкому ухудшению эффективности кодирования.

В то же время с появлением модификации ЭРГК, оперирующей группами символов постоянного (заранее заданного) размера [11], при сравнимой с ЭРГК эффективностью кодирования, становится возможной разработка варианта ЭРГК с динамическим моделированием (ДЭРГК), являющаяся *целью данной работы.* Для этого в данной работе предлагается выбрать размеры шестнадцати групп и не менять их в процессе кодирования, а перечень символов в группах не сохранять вместе со сжатым текстом, а определять динамически в соответствии с их частотами, вычисляемыми в скользящем окне.

В первом подразделе описывается предлагаемая методика. Во втором подразделе приводится используемая тестовая модель данных, а также метрика для оценивания степени статистической однородности сжимаемого текста. И, наконец, в третьем подразделе работы проводится сравнительный анализ эффективности ЭРГК и ДЭРГК, а также широко известного архиватора WinRar [12].

1. Описание предлагаемой модификации

Представим ДЭРГК в виде функции от исходного текста и параметров сжатия:

$$C = \text{ДЭРГК}(M, D, L, P), \quad (1)$$

где C - закодированный текст, M - исходный текст, D - размер скользящего окна для вычисления частотных характеристик текста, L - предварительно выбранный массив размеров супербукв $L = \{L_1, L_2, ..., L_S\}$, не меняющийся в процессе кодирования [11], S - количество супербукв, P - заданный период обновления таблиц супербукв.

Обозначим как N длину исходного текста M. Тогда алгоритм кодирования M_i символа текста (i-индекс символа, изменяющийся в диапазоне 1...N) будет выглядеть следующим образом:

1. Если $i-D > 0$, то уменьшается на единицу значение ячейки массива частот F с индексом M_{i-D}, где F - массив целых чисел с номерами ячеек от 0 до K-1, где K - размер исходного алфавита (в данной работе для простоты будем использовать K=256).

2. Если $i-1 > 0$, то увеличивается на единицу значение ячейки массива частот F с индексом M_{i-1}.

3. Если i mod P = 0 (i делится без остатка на P), то символы текста упорядочиваются в соответствии с убыванием их частот F и распределяются между S супербуквами в соответствии с размерами супербукв L (в данной работе S=16).

4. Символ Mi разделяется на префикс (номер супербуквы) и суффикс (порядковый номер символа внутри супербуквы). Префикс помещается в поток префиксов, а суффикс - в поток суффиксов.

Перед кодированием первого символа алфавита M_1 массив значений F заполняется единицами.

В качестве массива L в данной работе будем использовать рекомендованный в [11] массив {1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 4, 16, 32, 64, 128}.

Параметр D определяет, насколько быстро предлагаемый метод будет адаптироваться к изменению частотных характеристик текста.

Чем меньшим будет значение D, тем больше будет адаптивность метода к изменениям частотных характеристик текста, однако тем меньшей будет статистическая достоверность моделей, используемых при кодировании. Как показывает практика, изменение D в пределах от 1024 до 4096 способно, в зависимости от конкретного M, приводить к увеличению или уменьшению коэффициентов сжатия на 0.1...0.3%. Для модели тестовых данных, используемой в данной работе (см. подраздел 2), такие изменения являются несущественными, поэтому в данной работе будет использоваться D=2048.

Параметр P влияет одновременно на эффективность сжатия и скорость кодирования текста.

При P=1 обеспечивается наиболее высокий коэффициент сжатия текста. Однако при этом резко падает скорость кодирования, так как перед кодированием каждого символа текста (это всего лишь несколько простых операций [7]) необходимо сортировать символы по их частотам и обновлять таблицы супербукв. В работе рекомендуется значение P=512. Коэффициенты сжатия при этом уменьшаются на 0,1...0,2%, однако скорость кодирования возрастает в 512 раз и становится сравнимой со скоростью кодирования для ЭРГК со статическим частотным моделированием.

При кодировании статистически неоднородного M формируемый массив префиксов (номеров супербукв) так же может являться статистически неоднородным. В таком случае на следующей итерации к нему может применяться ДЭРГК. В случае же, если этот массив является статистически частотно однородным, то эффективнее будет применить для его кодирования ЭРКГ. На практике можно просто сжать массив префиксов обоими методами и выбрать лучший результат. В то же время для ускорения кодирования можно оценить статистическую однородность текста с помощью какой-либо метрики и на основании ее значения выбрать ЭРГК или ДЭРГК. Возможный вариант такой метрики рассматривается в следующем подразделе.

2. Модель тестовых данных

В данной работе используется следующая модель статистически частотно неоднородных данных:

$$M_i = [i\ K/N + \xi\ i/N], \quad (2)$$

где ξ - случайное число с Гауссовым законом распределения, нулевым математическим ожиданием и сдреднеквадратическим отклонением σ, [] - операция округления.

При этом, если получилось значение M_i, больше, чем K-1, то из него вычитается (K-1). Если значение M_i получилось меньше, чем 0, то к нему прибавляется (K-1).

Параметр σ обеспечивает предлагаемой модели высокую гибкость. Использование больших σ повышает статистическую однородность данных. Использование малых σ ее понижает. В данной работе в сравнительном анализе будем использовать σ в диапазоне от 3 до 30.

Множитель для σ сильно возрастает к концу текста, что приводит к увеличению статистической неоднородности этого текста.

При этом построение таблицы частот символов по всему тексту, сгенерированному с использованием (2), приводит к примерно одинаковым частотам для всех символов текста. Это делает неэффективными для кодирования такого текста все методы сжатия, подразумевающие статическое частотное моделирование.

Для оценивания статистической однородности текста будем использовать метрику частотной однородности текста (МЧОТ), основанную на расстоянии Кульбака-Лейбера [13]:

$$МЧОТ = 2 / (1 + R_1/R_2), \quad (3)$$

где $R_1 = -\sum_{i=1}^{N} p(M_i)\log_2 p(M_i)$, $p(M_i)$ - вероятность встречаемости символа M_i в тексте;

$R_2 = -\sum_{i=1}^{N} p_l(M_i)\log_2 p_l(M_i)$, $p_l(M_i)$ - вероятность встречаемости символа M_i, вычисленная для локального фрагмента текста M, предшествующего символу Mi (в данной работе для этого используется фрагмент длиной 4096 символов).

Значение предложенной метрики стремится к единице для статистически однородных текстов. Для частотно неоднородных тестов значение МЧОТ будет лежать на промежутке от 0 до 1, причем, чем ближе оно будет к 0, тем о большей неоднородности текста это свидетельствует.

3. Сравнительный анализ

Проведем сравнение эффективности сжатия тестовых данных, полученных в соответствии с (2) для методов ЭРГК, ДЭРГК и широко известного метода WinRar.

Будем использовать тестовые данные с σ = 3, 5, 10, 15, 20, 30 и N=1048576. Значение bps = 8/ КС, (где КС - коэффициент сжатия) и значения МЧОТ приведены в таблице 1.

Таблица 1

Эффективность сжатия тестовых данных

σ тестового текста	МЧОТ	bps		
		WinRar	ЭРГК	ДЭРГК
3	0,51	2,75	8,0	2,58
5	0,58	3,44	8,0	3,27
10	0,68	4,52	8,0	4,23
15	0,74	5,02	8,0	4,79
20	0,78	5,35	8,0	5,20
30	0,83	5,85	8,0	5,77

Из данных таблицы видно, что стандартный вариант ЭРГК для кодирования таких данных не эффективен. В то же время предложенный метод ДЭРГК обеспечивает тем больший выигрыш в степени сжатия по отношению к ЭРГК, чем меньше МЧОТ. Для низких МЧОТ этот выигрыш в объеме сжатых данных достигает 65%.

Следует отметить, что для задачи сжатия подобных данных ДЭРГК демонстрирует более высокую эффективность не только по отношению к ЭРГК, но и к широко известному архиватору WinRAR, выигрывая у него по bps от 1% до 6% в зависимости от МЧОТ (чем меньше МЧОТ, тем больше выигрывает ДЭРГК у WinRar).

Следует также отметить, что эффективность оценивания статистической однородности текста с помощью МЧОТ косвенно подтверждается тем, что с ростом значения σ (большие значения соответствуют более высокой статистической однородности генерируемых тестовых данных) значение МЧОТ стабильно увеличивается.

Дополнительный численный анализ показал, что ДЭРГК обеспечивает более высокие КС, чем ЭРГК для текстов с МЧОТ меньше или равным 0,98. При значениях МЧОТ 0,99 и выше более эффективным является статическое моделирование в сочетании со стандартным ЭРГК.

Заключение

Проведенные исследования показали более высокую эффективность предложенного метода ДЭРГК по отношению к ЭРГК и к известному архиватору WinRar при сжатии статистически неоднородных данных.

Предложена также метрика МЧОТ, основанная на расстоянии Кульбака-Лейбера и позволяющая

эффективно оценивать степень статистической однородности текста. Показано, что данная метрика может использоваться в качестве индикатора эффективности использования ДЭРГК.

В качестве дальнейших исследований можно проанализировать эффективность ДЭРГК для других моделей текстов.

Литература

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

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

3. Burrows, M. A block sorting lossless data compression algorithm. Technical Report 124: Digital Equipment Corporation [Text] / M. Burrows, D. Wheeler. – Systems Research Center – 1994. – 24 p.

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

5. Huffman, D. A. A method for the construction of minimum-redundancy codes [Text] / D. A. Huffman // Proceedings of Institute of Radio Engineering – September, 1952. – Vol. 40, N 9. – P. 1098-1101.

6. 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.

7. 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.

8. Кожемякина, Н. В. Сравнительный анализ эффективности методов сжатия данных при кодировании символов больших алфавитов [Текст] / Н. В. Кожемякина, Н. Н. Пономаренко, А.А. Зеленский // Системи обробки інформації. – 2015. – № 9 (134). – С. 74-78.

9. JPEG сжатие изображений с применением рекурсивного группового кодирования [Текст] / Н. В. Кожемякина, В. В. Лукин, Н. Н. Пономаренко, А. И. Мирошниченко // Радіоелектронні і комп'ютерні системи. – 2015. – № 3 (73). – С. 77-81.

10. Method of data compression for traffic monitoring [Text] / N. Kozhemiakina, V. Lukin, N. Ponomarenko, A. Akulynichev, K. Egiazarian, J. Astola // IEEE Second International Scientific-Practical Conference «Problems of Infocommunications Science and Technology (PIC S&T – 2015)». – 2015. – P. 153-156.

11. Пономаренко, Н. Н. Рекурсивное групповое кодирование с количеством и размерами групп, не зависящими от кодируемых данных [Текст] / Н. Н. Пономаренко, Н. В. Кожемякина // Радіоелектронні і комп'ютерні системи. – 2015. – № 2 (72). – С. 112-115.

12. The WinRar committee home page [Electronic resource]: Data compression programs, website. – Access mode: http://www.rarlab.com – Access date 05.06.2016. – Title by screen

13. Kullback, S. On information and sufficiency [Text] / S. Kullback, R. Leibler // The Annals of Mathematical Statistics. – 1951. – V. 22, N. 1. – P. 79-86.

References

1. Salomon, D. *Data Compression - The Complete Reference*. Springer-Verlag, 2004. 898 p.

2. Cleary, J., Witten, I. Data compression using adaptive coding and partial string matching. *IEEE Transactions on Communications.,* 1984, vol. 32, pp. 396-402.

3. Burrows, M., Wheeler, D. *A block sorting lossless data compression algorithm. Technical Report 124: Digital Equipment Corporation*. Systems Research Center Publ., 1994. 24 p.

4. Rissanen, J. Generalized kraft inequality and arithmetic coding. *IBM J. Res. Develop*, 1976, vol. 20, pp. 198-203.

5. Huffman, D. A. A method for the construction of minimum-redundancy codes. *Proceedings of Institute of Radio Engineering*, 1952, vol. 40, no. 9, pp. 1098-1101

6. Ponomarenko, N., Lukin, V., Egiazarian, K., Astola, J. Fast recursive coding based on grouping of Symbols. *Telecommunications and Radio Engineering*, 2009, vol. 68, no. 20, pp. 1857-1863.

7. Ponomarenko, N. N., Kozhemyakina, N. V., Lukin, V. V. Metod entropiinogo rekursivnogo gruppovogo kodirovaniya [Method of entropy recursive group coding]. *Radioelektronni i komp'yuterni sistemi*, 2014, vol. 3, no. 67, pp. 20-26.

8. Kozhemyakina, N. V., Ponomarenko, N. N., Zelenskii, A. A. Sravnitel'nyi analiz effektivnosti metodov szhatiya dannykh pri kodirovanii simvolov bol'shikh alfavitov [Comparative analysis of data compression methods for encoding of symbols of large alphabets]. *Sistemi obrobki informatsiї*, 2015, vol. 9, no. 134, pp. 74-78.

9. Kozhemyakina, N. V., Lukin, V. V., Ponomarenko, N. N., Miroshnichenko, A. I. JPEG szhatie izobrazhenii s primeneniem rekursivnogo gruppovogo kodirovaniya [JPEG image compression using recursive group coding]. *Radioelektronni i komp'yuterni sistemi*, 2015, vol. 3, no. 6,3 pp. 77-81.

10. Kozhemiakina, N., Lukin, V., Ponomarenko, N., Akulynichev A., Egiazarian, K., Astola, J. Method of data compression for traffic monitoring. *IEEE Second International Scientific-Practical Conference «Problems of Infocommunications Science and Technology (PIC S&T – 2015)»,* Kharkov, Ukraine, 2015, pp. 153-156.

11. Ponomarenko, N. N., Kozhemyakina, N. V. Rekursivnoe gruppovoe kodirovanie s kolichestvom i razmerami grupp, ne zavisyashchimi ot kodiruemykh dannykh [Recursive group coding with fixed number

and sizes of groups]. *Radioelektronni i komp'yuterni sistemi*, 2015, vol. 2, no. 72, pp. 112-115.

12. *The WinRar committee home page: Data compression programs.* Available at: http://www.rarlab.com (accessed 05.06.2016).

13. Kullback, S., Leibler, R. On information and sufficiency. *The Annals of Mathematical Statistics,* 1951, vol. .22, no. 1, pp. 79-86.

Поступила в редакцию 20.10.2016, рассмотрена на редколлегии 09.12.2016

РЕКУРСИВНЕ ГРУПОВЕ КОДУВАННЯ З РЕКУРСИВНИМ ЧАСТОТНИМ МОДЕЛЮВАННЯМ

Н. В. Кожемякіна, М. М. Пономаренко

Розглянуто задачу ентропійного кодування даних з метою усунення в них статистичної надмірності на основі рекурсивного групового кодування. Рекурсивне групове кодування є більш швидкою і в ряді випадків більш ефективною альтернативою таким методам, як арифметичне кодування і кодування Хаффмана, за рахунок здатності ефективно кодувати символи надвеликих алфавітів. Однак його застосування до цих пір обмежувалося статичним частотним моделюванням, при якому для всього тексту обчислюються частоти символів, а інформація про розміри та склад груп зберігається в файл разом зі стисненим текстом. У даній роботі вперше пропонується модифікація рекурсивного групового кодування, що дозволяє ефективно стискати статистично неоднорідні за частотними характеристиками тексти. Розглянуто модель формування тестових наборів даних, запропоновано метрику для оцінювання статистичної неоднорідності даних. Показано, що для розглянутих тестових даних запропонована модифікація забезпечує до 65% менший обсяг стислих даних, ніж стандартний варіант рекурсивного групового кодування.

Ключові слова: рекурсивне групове кодування, ентропійне кодування, арифметичне кодування, кодування Хафмана.

RECURSIVE GROUP CODING WITH DYNAMIC FREQUENCY MODELING

N. V. Kozhemiakina, N. N. Ponomarenko

Task of entropy group coding of data for reduce of its statistical redundancy on base of recursive group coding is considered. Recursive group coding provides effective coding of symbols of large alphabets. It is fast and in some cases more effective alternative of such methods as arithmetical coding and Huffman coding. However applications of recursive group coding are restricted by usage of statistical frequency modeling. For such modeling frequencies (or probabilities) of symbols are calculated for entire text as well as information about size sand contents of groups are stored in compressed file. In this work a modification of recursive group coding are proposed which is able effectively compress texts statistically heterogeneous by frequency characteristics. A model for synthesizing of test texts is considered. Also a new metric for estimates of uniformity of a given text is proposed. It is shown that for considered test texts the proposed modifications of recursive group coding provide up to 65% less compressed data size than conventional method.

Key words*:* recursive group coding, entropy coding, arithmetical coding, Huffman coding.

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

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

Kozhemiakina Nadejda Vladimirovna – assistant, Department of Transmitters, Receivers and Signal Processing, National Aerospace University named after N. Ye. Zhykovsky «KhAI», Kharkov, Ukraine, e-mail: nadejda_kozickaya@mail.ru.

Ponomarenko Nikolay Nikolaevich – Doctor of Technical Sciences, Associate Professor, Professor of Department of Transmitters, Receivers and Signal Processing, National Aerospace University named after N. Ye. Zhykovsky «KhAI», Kharkov, Ukraine, e-mail: nikolay@ponomarenko.info.