УДК 621.391

Н.В. Кожемякина, Н.Н. Пономаренко

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

ЭНТРОПИЙНОЕ РЕКУРСИВНОЕ ГРУППОВОЕ КОДИРОВАНИЕ ДЛЯ ДВУХБАЙТНЫХ АЛФАВИТОВ

. В данной работе предлагается модификация энтропийного рекурсивного группового кодирования (ЭРГК), которая за счет использования динамического частотного моделирования и ЭРГК с фиксированными размерами групп позволяет на первой итерации кодировать двухбайтные символы. Предлагается модель формирования тестовых данных с двухбайтным алфавитом, позволяющих подтвердить эффективность данной модификации ЭРГК. Показано, что предложенная модификация ЭРГК для таких данных обеспечивает более высокую эффективность сжатия не только, чем арифметическое кодирование и кодирование Хаффмана, но и чем эффективные высокоуровневые методы сжатия, такие как WinRar и PAQ8.

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

Введение

Методы устранения статистической избыточности в данных, такие как кодирование Хаффмана (КФ) [1] или арифметическое кодирование (АК) [2] практически не находят самостоятельного применения, однако входят в состав большинства методов сжатия изображений (таких, как стандарт JPEG [3]) и видео (таких, как стандарт H.265 [4]), а также универсальных архиваторов для сжатия данных (таких, как WinRar [5] или PAQ8 [6]). Можно сказать, что КФ, АР и подобные им методы используются в высокоуровневых схемах сжатия в качестве "элементарных кирпичиков" для энтропийного кодирования информации с целью уменьшения ее объема.

Энтропийное рекурсивное групповое кодирование (ЭРГК) является новой быстрой и эффективной альтернативой таким методам, как АК или КХ, обеспечивая существенно более высокие степени сжатия для текстов со сверхбольшими алфавитами. К таким текстам относятся, например, кодируемые мультимедийные данные (изображения, звук и видео) при их сжатии с потерями. При этом ЭРГК применяется к символам небольшой размерности, но за счет своей рекурсивности и объединения в процессе кодирования пар соседних символов, на каждой последующей итерации размерность символов удваивается. Однако из-за необходимости сохранять в сжатых данных списки групп символов, все предложенные ранее методы работают с однобайтными символами, что в ряде случаев может приводить к падению эффективности кодирования данных

Метод ЭРГК [7; 8] был разработан в качестве вычислительно простой, быстрой и эффективной альтернативы КХ и АК. При кодировании символов однобайтных алфавитов (когда корреляция между соседними байтами входного потока данных практически отсутствует) ЭРГК обеспечивает степени сжатия, сравнимые с КХ и АК. В случае же наличия корреляции между соседними байтами ЭРГК способно эффективно ее учитывать и обеспечивать существенно более высокую степень сжатия, чем КХ и АК. Кроме того, недостатком кодирования Хаффмана является недостаточная эффективность кодирования символов, для хранения которых может быть достаточно менее одного бита памяти, а недостатком арифметического кодирования является использование при кодировании операций умножения и деления, что негативно сказывается на скорости кодирования, значительно снижая быстродействие.

Более высокая степень сжатия достигается использованием рекурсии. В исходном тексте подсчитываются частоты встречаемости символов, а затем происходит их сортировка по возрастанию. Далее происходит формирование супербукв (групп символов) объединением в супербукву символов алфавита с близкими частотами их встречаемости в тексте. Через объединение символов алфавита в супербукву увеличивается длина кода для этих символов. При этом супербуква может быть образована только в том случае, если это увеличение длины кода не превышает заданного допустимого порога. Затем выделяют префиксы (номера групп) и суффиксы (номер символа внутри группы). Суффиксы нумеруются и вместе с таблицами супербукв хранятся в файле, который сжимается (для обеспечения в дальнейшем проведения процедуры декодирования). При этом к префиксам не применяется кодирование, а происходит объединение рядом расположенных префиксов

(2 и более префикса), и образование нового текста, который в несколько раз меньше исходного. Далее метод применяется к образованному более короткому тексту, то есть действия рекурсивно повторяются, пока размер этого текста не станет меньше заданного порога [7; 8].

Хорошо исследованным для ЭРГК является кодирование с интерпретацией символов алфавита, как однобайтных (длина алфавита – 256 символов) [9]. Для этого варианта возможно использование статического частотного моделирования, предусматривающего сохранение в сжатых данных перечней символов, входящих в ту или иную группу (супербукву). Для двухбайтных алфавитов (длина алфавита – 65536 символов) подобное сохранение списка символов в сжатых данных в большинстве ситуаций делает сжатие неэффективным (закодированные данные занимают больший объем, чем исходные).

В работе [10] было показано, что в ЭРГК при кодировании можно использовать фиксированные размеры и количество групп символов без существенного уменьшения степени сжатия данных. Это позволило в работе [11] предложить динамический вариант ЭРГК (ДЭРГК), который способен эффективно сжимать текст с нестационарными (меняющимися для разных участков текста) частотными характеристиками символов. В [11] с помощью специально разработанной модели текста показано, что ДЭРГК способен эффективно кодировать текст даже в ситуации, если в целом по всему тексту частоты всех символов примерно одинаковы, но существенно отличаются на отдельных участках (ЭРГК не способен сжимать такой текст).

Целью данной работы является разработка модификации ДЭРГК для эффективного кодирования текстов двухбайтных алфавитов. Для этого в данной работе предлагается на первой итерации использовать ДЭРГК с двухбайтным алфавитом, фиксированными размерами групп и количеством групп символов, равным 16. Это обеспечит для второй итерации длину алфавита в 256 символов и позволит на всех последующих итерациях использовать стандартный ЭРГК или ДЭРГК для однобайтного алфавита.

Предлагаемая модификация описывается в первом подразделе работы, в то время как второй подраздел посвящен численному анализу.

1. Описание предлагаемой модификации

Ключевая идея предлагаемого метода состоит в использовании при кодировании разбиения алфавита на 16 групп (супербукв). Таким образом, для всего разнообразия символов исходного алфавита будет иметься всего 16 разных префиксов (номеров супербукв). Их попарное объединение даст 16^2 вариантов, то есть 256 символов нового алфавита. Тогда на второй итерации к новому тексту, образованному объединением соседних префиксов, можно будет применять ЭРГК для однобайтного алфавита.

Требуемые 16 групп символов должны в общей сложности включать в себя 65536 символов, при этом размер каждой группы должен быть кратным степени двойки [8]. Поэтому в данной работе предлагается использовать не изменяющийся в процессе кодирования список групп символов с размерами L={2, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096, 8192, 16384, 32768). В сумме размеры всех этих групп составляют требуемые 65536.

Предлагаемая модификация ДЭРГК для двухбайтного алфавита (ДЭРГК2) состоит из следующих шагов:

1. Исходный текст M сжимается ДЭРГК с формированием кода C в соответствии с выражением C=ДЭРГК(M, D, L, P) [11], где размер скользящего окна для вычисления частот символов D = 2048, период обновления таблиц супербукв P = 512 (выбор значений параметров D и P пояснен в [11]).

2. Поток суффиксов из кода C сохраняется в сжатых данных без изменений.

3. Поток объединенных пар префиксов из кода C подвергается сжатию стандартным ЭРГК для однобайтного алфавита, если значение метрики МЧОТ [11] для этого потока превышает 0,98. В противном случае этот поток подвергается сжатию ДЭРГК для однобайтного алфавита.

Данная модификация может быть неэффективной в случае, если вероятность встречаемости одного из символов двухбайтного алфавита в несколько раз превышает вероятности встречаемости всех остальных символов. Для этого частного случая можно использовать несколько избыточную по общему количеству символов таблицу размеров групп L={1, 4, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096, 8192, 16384, 32768) или же одну из модификаций ЭРГК для однобайтных алфавитов.

Отметим, что из-за необходимости периодически сортировать массив частот встречаемости 65536 символов и распределять их по группам, ДЭРГК2 будет работать существенно медленнее, чем ЭРГК.

2. Сравнительный анализ

В данной работе используется та же модель статистически частотно неоднородных данных, что и в [11]:

$$M_i = [i\,K/N + \xi\, i/N], \qquad (1)$$

где ξ – случайное число с Гауссовым законом распределения, нулевым математическим ожиданием и среднеквадратическим отклонением σ; i - номер буквы текста; K=65536 (размер алфавита); N – длина текста; [] –- операция округления.

Отличие в тестовых данных, используемых в

данной работе от [11], состоит в том, что, кроме большего значения K, для полученных по (1) данных выполняется еще и операция подстановки:

$$M_i = S(M_i), \quad (2)$$

где S – массив чисел от 0 до K–1, отсортированный в случайном порядке.

Операция (2) затрудняет кодирование для некоторых высокоуровневых методов сжатия, которые могут учитывать близость вероятностей встречаемости в тексте для (1) символов с близкими численными значениями.

Проведем сравнительный анализ эффективности сжатия тестовых данных методами ЭРГК (P=512, D=2048), АК, КХ (все – со статическим частотным моделированием для однобайтных алфавитов), ДЭРГК и предложенным ДЭРГК2.

Кроме того, включим в анализ и высокоуровневые методы сжатия (широко распространенный архиватор WinRar и архиватор PAQ8, обеспечивающий на данный момент одни из наилучших степеней сжатия для большинства типов данных).

Будем использовать тестовые данные с σ = 30, 50, 100, 200, 400 и N=8388608 (2^{23}). Количество бит на один символ сжатого текста (bps = 8/ КС, где КС – коэффициент сжатия) приведены в табл. 1.

Кодирование для PAQ8 осуществлялось с ключом "-8", а для WinRar - с ключом "-m5". Оба ключа переключают соответствующие архиваторы в режим медленного кодирования, обеспечивающего максимальную степень сжатия.

Таблица 1

Эффективность сжатия тестовых данных, bps

σ	PAQ8	WinRar	ЭРГК, АК, КХ	ДЭРГК	ДЭРГК2
30	6,1	8,2	16	13,0	5,7
50	6,9	9,5	16	13,9	6,0
100	7.9	11,2	16	13,9	6,8
200	8,9	12,8	16	15,4	7,9
400	9,9	14,1	16	15,7	9,2

Из данных таблицы видно, что статические методы ЭРГК, АК и КХ для кодирования таких данных неэффективны, так как вероятности встречаемости всех символов для модели (1) примерно одинаковы, если брать весь текст в целом.

В то же время метод ДЭРГК за счет динамического кодирования способен сжимать тестовые данные, однако не достаточно эффективно, так как он работает с однобайтными символами.

Предложенная модификация ДЭРГК2 обеспечивает наибольшие КС для тестовых данных среди всех рассмотренных методов. Более того, для данной модели данных она превосходит эффективные высокоуровневые методы сжатия, такие как PAQ8, в некоторых случаях обеспечивая на 15% меньший объем сжатых данных. При этом метод PAQ8 является более вычислительно сложным и намного более медленным, чем ДЭРГК2.

Выводы

Предложенная в работе модификация ДЭРГК2 метода ЭРГК позволяет эффективно кодировать данные двухбайтных алфавитов для текстов, неоднородных по вероятностям встречающихся символов. Проведенный анализ показал, что ДЭРГК2 для кодирования таких данных превосходит даже сложные высокоуровневые методы сжатия, такие как WinRar или PAQ8.

Основную проблему при кодировании с помощью ДЭРГК2 составляет необходимость периодической сортировки таблиц частот встречаемости символов, что существенно замедляет процесс кодирования. Поэтому в качестве дальнейших исследований актуальной является разработка быстрых алгоритмов сортировки таблиц частот символов и обновления списков групп для метода ДЭРГК2.

Список литературы

1. 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.

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

3. Wallace, G. The JPEG Still Picture Compression Standard [Text] / G. Wallace // Comm. of the ACM. – 1991. – Vol. 34. – P. 30-44.

4. Recommendation ITU-T H.265: High efficiency video coding. – Geneva, Switzerland. – 610 p. [Electronic resource] Access mode: https://www.itu.int/rec/T-REC-H.265-201504-I/en.

5. The WinRar committee home page [Electronic resource] // Data compression programs, website. – Access mode: http://www.rarlab.com – Access date 25.02.2017. – Title by screen.

6. Mahoney, M The PAQ [Electronic resource] // Data compression Programs. – Access mode: http://mattmahoney.net/dc/paq.html – Access date 25.02.2017. – Title by screen.

7. 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.

8. 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.

9. 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. – Х.: ХУПС, 2015. – № 9 (134). – С. 74-78.

10. 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.

11. Кожемякина, Н. В.Рекурсивное групповое кодирования с динамическим частотным моделированием [Текст] / Н. В. Кожемякина, Н. Н. Пономаренко // Радіоелектронні і комп'ютерні системи. – 2016. – № 4 (78). – С. 22-26.

Поступила в редколлегию19.05.2017

Рецензент: д-р техн. наук проф. В.К. Волосюк, Национальный аэрокосмический университет им. Н. Е. Жуковского «ХАИ», Харьков.

ЕНТРОПІЙНЕ РЕКУРСИВНЕ ГРУПОВЕ КОДУВАННЯ ДЛЯ ДВОБАЙТОВИХ АЛФАВІТІВ

Н.В. Кожемякіна, М.М. Пономаренко

В даній роботі пропонується модифікація ентропійного рекурсивного групового кодування (ЕРГК), яка за рахунок використання динамічного частотного моделювання і ЕРГК з фіксованими розмірами груп дозволяє на першій ітерації кодувати двухбайтні символи. Пропонується модель формування тестових даних з двухбайтним алфавітом, що дозволяють підтвердити ефективність даної модифікації ЕРГК. Показано, що запропонована модифікація ЕРГК для таких даних забезпечує більш високу ефективність стиснення не тільки, ніж арифметичне кодування і кодування Хафмана, але й ніж ефективні високорівневі методи стиснення, такі як WinRar і PAQ8.

Ключові слова: *рекурсивне групове кодування, ентропійне кодування, арифметичне кодування, кодування Хафмана, динамічне частотне моделювання.*

ENTROPY RECURSIVE GROUP CODING FOR DOUBLE-BYTE ALPHABETS

N. Kozhemiakina, N. Ponomarenko

In the paper a modification of entropy recursive group coding (ERGC) that by usage of dynamic frequency modeling and fixed sizes of group allows in the first iteration to code double-byte symbols is proposed. A model of test text with double-byte symbols allowing to approve effectiveness of the modification of ERGC is described. It is shown that the proposed modification provides larger compression ratios not only than arithmetical coding and Huffman coding but also than such high-level compression methods as WinRar and PAQ8.

Keywords*: recursive group coding, entropy coding, arithmetical coding, Huffman coding, dynamic frequency modeling.*