УДК 621.391

В. А. ЛУЖЕЦЬКИЙ, Л.А. САВИЦЬКА

Вінницький національний технічний університет, Вінниця

АДАПТИВНИЙ МЕТОД УЩІЛЬНЕННЯ ДАНИХ ОДНИМ ПРОХОДОМ З РІВНОМІРНИМ РОЗБИТТЯМ НА БЛОКИ

Анотація. Розглянуто метод ущільнення даних на основі лінійної форми Фібоначчі, який передбачає використання двох моделей джерела даних, чотирьох правил кодування даних і функції оптимізації (адаптації) та забезпечує підвищення коефіцієнта ущільнення за рахунок адаптації виконуваних перетворень до конкретного змісту ущільнюваних даних.

Ключові слова: метод ущільнення, лінійна форма Фібоначчі, адаптація, моделі джерела даних, кодування даних, декодування даних.

Аннотация. Рассмотрен метод сжатия даних на основе линейной формы Фибоначчи, предусматривающий использование двух моделей источника данных, четырех правил кодирования данных и функции оптимизации (адаптации), обеспечивающий повышение коэффициента сжатия за счет адаптации выполняемых преобразований к конкретному содержанию сжимаемых данных.

Ключевые слова: метод сжатия, линейная форма Фибоначчи, адаптация, модели источника данных, кодирование данных, декодирование данных.

Annotation. There was considered the method of data compaction which is based on Fibonacci linear form. This method involves the use of two models of data sources,four rules of data coding and the function of optimization/adaptation. It ensures increased the compaction coefficient due to adaptation of executable transformation to specific content by data compaction.

Key words: Method of compaction; Fibonacci linear form; adaptation; models of data sources; coding data; decoding data

Вступ

Існуючі методи й алгоритми ущільнення даних, в основному, враховують статистику символів у повідомленні [1–3] або базуються на побудові словника [4, 5]. І хоча на практиці широко використовуються архіватори, створені на основі саме цих методів, пошуки нових підходів до ущільнення даних продовжуються.

Одним із цікавих підходів є пропозиція використовувати для ущільнення даних оптимізуючі властивості чисел Фібоначчі [6, 7]. Суть підходу полягає в тому, що в процесі ущільнення інформації блок цифрових даних будь-якої довжини розглядається як надвелике ціле додатне число, що представляється набором із трьох невеликих чисел. Таке представлення чисел, називається лінійною формою Фібоначчі.

Актуальність

Теоретичні дослідження і практика застосування архіваторів показали, що не існує універсального неадаптивного методу ущільнення, що забезпечував би однаковий коефіцієнт ущільнення для різних типів даних. Тому наукові дослідження найчастіше спрямовані на створення ефективних методів ущільнення певних типів даних. Однак дані навіть одного типу, з погляду ущільнення, мають різні властивості і характеристики. З огляду на це, останнім часом прагнуть до створення адаптивних алгоритмів ущільнення даних [8, 9]. Запропоновані в роботах [6, 7] методи ущільнення даних на основі лінійної форми Фібоначчі є неадаптивними і тому не можуть забезпечити однаково високий коефіцієнт ущільнення для різних типів даних.

Мета дослідження

Метою статті є підвищення ефективності ущільнення даних на основі лінійної форми Фібоначчі за рахунок адаптації виконуваних перетворень до конкретного змісту ущільнюваних даних.

Постановка задач

Одним з найважливіших положень теорії ущільнення інформації є висловлена в [10] ідея поділу процесу ущільнення на дві процедури: моделювання і кодування. Моделювання визначає характеристики джерела даних, що ущільнюються, а кодування перетворює символи у послідовність бітів відповідно до отриманих характеристик.

У роботі [11] запропоновано таку узагальнену модель процесу адаптивного ущільнення:

$$\mathbf{C}_A = \left\{ \mathbf{P}, \mathbf{A}, \mathbf{M}, \mathbf{C}, \mathbf{D}, \mathbf{P}_M, \mathbf{P}_C, \mathbf{P}^*, S, f \right\}$$

де P - вихідна послідовність символів алфавіту $\mathbf{A} = \{0,1\}$; $\mathbf{M} = \{M_i\}$ - множина правил моделювання джерела даних; $\mathbf{C} = \{C_j\}$ - множина правил кодування даних; $\mathbf{D} = \{D_j\}$ - множина правил декодування даних; $\mathbf{P}_M = \{\mathbf{P}_{M_i}\}$ - множина послідовностей, що є результатом моделювання; $\mathbf{P}_C = \{\mathbf{P}_{C_{ij}}\}$ -

множина послідовностей, що є результатом кодування; $\mathbf{P}^*$ - послідовність ущільнених даних; $\boldsymbol{S}$ - правило формування структури послідовності $\mathbf{P}^*$; f - функція оптимізації.

Саме наявність множини правил моделювання джерела даних і множини правил кодування даних забезпечують можливість адаптації до конкретного змісту ущільнюваних даних. При цьому правило вибору з множини кодованих послідовностей $\mathbf{P}_C$ єдиної послідовності $\mathbf{P}^*$, що відповідає найбільшому коефіцієнту ущільнення, описується функцією оптимізації f.

Враховуючи сказане, в даній статті розв’язуються такі задачі:

– визначення множини правил моделювання джерела даних і множини правил кодування даних;

– розробка адаптивного методу ущільнення даних на основі лінійної форми Фібоначчі.

Розв’язання задач

Пропонується метод ущільнення, який полягає в тому, що вихідні дані розбиваються на блоки однакової довжини, кожен з яких кодується незалежно за чотирма правилами і з чотирьох результатів кодування вибирається той результат, структура блоку якого має найменшу довжину.

Множина правил моделювання джерела даних складається із двох правил $\mathbf{M} = \{m(a,0), m(1)\}$. Перше правило означає, що вихідні дані розбиваються на блоки довжини 2^a розрядів. При цьому кожен блок даних розглядається або як набір символів 0 і 1, або як число *N*, що обчислюється за формулою:

$$N = \sum_{i=0}^{2^a-1} c_i 2^i ,$$

де c_i - *i*-й символ блоку.

Друге правило моделювання джерела даних полягає в записі символу "1” в самий старший розряд блоку, якщо він містить нуль. Тобто блок вигляду 0xx...x заміняється на блок вигляду 1xx...x, де x означає будь-який символ (1 або 0).

Множину правил кодування вибрано виходячи з таких міркувань.

Перетворення, що використовуються для ущільнення, залежно від змісту блоку можуть або скоротити його розрядність, або не змінити, або збільшити. У двох останніх випадках доцільно залишити блок неперетвореним, дописавши до нього відповідну ознаку. Таке правило кодування будемо позначати $C_0(0,0)$.

Якщо блок має вигляд:

$$\underbrace{\underbrace{0\,0\ldots0}_{q}\;1\mathrm{x}\;\mathrm{x}\ldots\mathrm{x}}_{2^a},$$

то його можна перетворити відкиданням q старших нулів до вигляду:

$$\underbrace{1\mathrm{x}\;\mathrm{x}\ldots\mathrm{x}}_{2^a-q}\;\underbrace{\|\;q}_{a}$$

за умову $q > a+1$. Таке правило кодування будемо позначати $C_1(0,1)$.

Правило кодування $C_2(1,0)$ полягає в безпосередньому перетворенні числа, що відповідає коду блоку, в лінійну форму Фібоначчі. В цій формі будь-яке ціле додатне число представляється у вигляді [6]:

$$N = a_0 F(j) + b_0 F(j+1), \qquad (1)$$

де a_0, b_0 – цілі додатні числа (координати представлення); j – ціле додатне число (індекс представлення); $F(j)$ – j-е число Фібоначчі.

Правило кодування $C_3(1,1)$ передбачає запис одиниці в самий старший розряд блоку вигляду 0xx...x, тобто реалізується правило моделювання *m*(1), а потім виконується перетворення в лінійну форму Фібоначчі.

Таким чином, маємо чотири правила кодування $\mathbf{C} = \{C_0(0,0), C_1(0,1), \ C_2(1,0), C_3(1,1)\}$, які породжують відповідні структури кодованих блоків (табл.1). Причому кожне із правил передбачає формування ознак структури p_1 і p_2. Тут l_{a_0} - розрядність коду числа a_0; l_{b_0} - розрядність коду числа b_0.

Таблиця 1 – Правила кодування і структури блоків

Позначення правила кодування	Ознаки		Позначення структури блоку	Вигляд структури блоку
	p_2	p_1		
$C_0(0,0)$	0	0	STR0	$\underbrace{\text{x x}\ldots\text{x}}_{2^a} \| 0 \| 0$
$C_1(0,1)$	0	1	STR1	$\underbrace{1\text{x x}\ldots\text{x}}_{2^a-q} \underbrace{\| \quad q}_{a} \| 0 \| 1$
$C_2(1,0)$	1	0	STR2	$b_0 \| a_0 \| j \| l_{b_0} \| l_{a_0} \| 1 \| 0$
$C_3(1,1)$	1	1	STR3	$b_0 \| a_0 \| j \| l_{b_0} \| l_{a_0} \| 1 \| 1$

Оскільки блоки перетвореної послідовності даних, у загальному випадку, будуть мати різну структуру, то результатом ущільнення є неоднорідна послідовність. Виходячи із цього, функція оптимізації, що описує правило вибору однієї з чотирьох структур перетвореного блоку, має вигляд:

$$f_{\text{бл}}^{\text{нп}} = \min\{l_0, l_1, l_2, l_3\},$$

де l_0, l_1, l_2, l_3 - довжина блоку структури STR0, STR1, STR2, STR3, відповідно.

У загальному випадку, при розбитті вихідних даних на блоки довжини 2^a один блок може мати довжину менше 2^a. Тому необхідно або вказати довжину вихідних даних L і потім обчислити довжину меншого блоку, або вказати довжину l_K цього блоку. Другий варіант вимагає меншої розрядності коду, тому вибираємо його. Таким чином, додаткова інформація містить у собі значення розрядності 2^a і l_K.

Отже маємо таке правило формування структури послідовності $\mathbf{P}^*$, яке представляється у вигляді кортежу:

$$S = \left\{\text{Бл}^* K \| \ldots \text{Бл}^* 2 \| \text{Бл}^* 1 \| l_K \| 2^a\right\}.$$

Розрядність кодованих блоків $\text{Бл}^* i$ ($i = 1 \div K$) різна. Розрядність вихідних блоків може досягати значення 2^{16}, тому розрядність полів l_K і 2^a дорівнює 16.

З урахуванням вищесказаного, реалізація адаптивного методу ущільнення даних одним проходом передбачає виконання таких обчислень і перетворень.

Спочатку вводиться послідовність даних **P**, що підлягають ущільненню, і визначається довжина *L* цієї послідовності. Потім вводиться значення параметру *a* і обчислюється кількість блоків $K = \left\lceil L \big/ 2^a \right\rceil$, на які буде розбиватися послідовність **P.** Потім обчислюється довжина *K*-го блоку $l_K = L - (K-1)2^a$. Після цього створюється початкова послідовність ущільнених даних $\mathbf{P}^*$, що складається із двох елементів: l_K і 2^a. Усі описані дії стосуються ініціалізації процесу ущільнення.

Безпосередньо процес ущільнення складається з K циклів. У кожному з них зчитується черговий блок вихідних даних Бл*i*, над яким виконуються перетворення згідно правил кодування $C_0(0,0)$, $C_1(0,1)$, $C_2(1,0)$, $C_3(1,1)$ і створюються структури STR0, STR1, STR2, STR3. Для кожної структури визначається її довжина та відповідно до правила оптимізації $\min\{l_0, l_1, l_2, l_3\}$ вибирається структура найменшої довжини, якою доповнюється послідовність ущільнених даних $\mathbf{P}^*$.

Процес відновлення даних визначається структурою послідовності ущільнених даних $S = \left\{ \text{Бл}^* K \,\|\, \ldots \text{Бл}^* 2 \,\|\, \text{Бл}^* 1 \,\|\, l_K \,\|\, 2^a \right\}$ і множиною правил декодування $\mathbf{D} = \left\{ D_0(0,0), D_1(0,1), D_2(1,0), D_3(1,1) \right\}$.

Розглянемо правила декодування даних для кожної з описаних вище структур.

Вибір правила декодування для заданого блоку здійснюється на основі аналізу ознак p_1 і p_2 структури блоку. Після вибору правила декодування розряди ознак відкидаються і подальші дії виконуються над основною частиною структури блоку.

Якщо p_1=0 і p_2=0, то це означає, що блок має структуру STR0, до якої застосовується правило декодування $D_0(0,0)$. Дане правило полягає в тому, що основна частина структури блоку залишається незмінною.

Якщо p_1=1 і p_2=0, то до блоку структури STR1 необхідно застосувати правило $D_1(0,1)$. Воно полягає у виконанні таких дій. В основній частині структури

$$\underbrace{1\text{x}\,\text{x}\ldots\text{x}}_{2^a-q} \,\|\, \underbrace{q}_{a}$$

відкинути a молодших розрядів і до частини, що залишилася, дописати q старших нулів. У результаті цього буде отримано блок вигляду:

$$\underbrace{\underbrace{0\,0\ldots0}_{q}\ 1\text{x}\,\text{x}\ldots\text{x}}_{2^a}.$$

Якщо p_1=0 і p_2=1, то до блоку структури STR2 застосовується правило $D_2(1,0)$. За цим правилом, виходячи з чисел j, a_0 і b_0, виконується зворотне перетворення Фібоначчі, тобто обчислюється за формулою (1) ціле додатне число N, що відповідає даній лінійній формі Фібоначчі. Після цього визначається розрядність l^* двійкового коду числа N і обчислюється різниця $\Delta = 2^a - l^*$. Дана різниця показує, скільки нулів необхідно дописати в старші розряди коду числа N.

Якщо p_1=1 і p_2=1, то до блоку структури STR3 застосовується правило $D_3(1,1)$. Воно передбачає обчислення числа за заданою лінійною формою Фібоначчі й заміну самої старшої одиниці в коді цього числа на нуль.

Відновлення вихідних даних з ущільнених передбачає виконання таких обчислень та перетворень.

Спочатку вводиться послідовність даних $\mathbf{P}^*$, які підлягають відновленню. Потім зчитуються 16 наймолодших розрядів, що є кодом числа 2^a. Після цього дані розряди видаляються з послідовності $\mathbf{P}^*$. На наступному кроці також зчитуються 16 наймолодших розрядів, що є кодом числа l_K. Далі створюється порожня послідовність **P.** Усі ці дії спрямовані на ініціалізацію процесу відновлення.

Безпосередньо процес відновлення складається з K циклів. На початку кожного циклу зчитується два наймолодші розряди послідовності $\mathbf{P}^*$, що є кодом ознаки структури. Відповідно до цього коду реалізується певне правило декодування. Результатом є черговий блок відновлених даних, що додається в послідовність **P**. Структура, з якої отриманий цей блок, видаляється з послідовності $\mathbf{P}^*$.

Відомі [6,7] методи ущільнення даних на основі лінійної форми Фібоначчі передбачають тільки одне правило кодування $C_2(1,0)$ для кожного блоку вхідних даних. При цьому можливі такі випадки:

структура STR2 перетвореного блоку має довжину, що дорівнює довжині вихідного блоку;

структура STR2 перетвореного блоку має довжину, що менша за довжину вихідного блоку;

структура STR2 перетвореного блоку має довжину, що більша за довжину вихідного блоку.

У першому випадку нема ефекту ущільнення, у другому – є, а у третьому випадку відбувається збільшення розрядності даних. Виходячи з цього коефіцієнт ущільнення даних визначається за формулою:

$$S_{\text{ущ}}^{\text{на}} = \frac{L}{w_1 2^a + l_2 + l_3},$$

де $l_2 = \sum_{i=1}^{w_2} l_i$ – сумарна довжина перетворених блоків, що відповідають випадку 2; $l_3 = \sum_{j=1}^{w_3} l_j$ –

сумарна довжина перетворених блоків, що відповідають випадку 3; w_1, w_2, w_3 - кількість перетворених блоків, що відповідають випадкам 1, 2 і 3

$$w_1 + w_2 + w_3 = K .$$

У разі застосування запропонованого адаптивного методу ущільнення випадок 3 виключається, а замість нього буде випадок 1. При цьому коефіцієнт ущільнення даних буде визначатися за формулою:

$$S_{\text{ущ}}^{\text{a}} = \frac{L}{w_1 2^a + l_2},$$

де $w_1 = K - w_2$.

Експериментальні дослідження запропонованого адаптивного методу ущільнення показали, що він забезпечує підвищення коефіцієнта ущільнення для різних типів файлів від 2 до 6 разів порівняно з неадаптивним методом ущільнення на основі лінійної форми Фібоначчі.

Висновки

1. Запропоновано новий метод ущільнення даних на основі лінійної форми Фібоначчі, який передбачає використання двох моделей джерела даних, чотирьох правил кодування даних і функції оптимізації (адаптації). Вибір на основі функції оптимізації моделі джерела даних і правила кодування даних, які спільно забезпечують найбільший коефіцієнт ущільнення, сприяє адаптації виконуваних перетворень до конкретного змісту ущільнюваних даних.

2. Адаптація забезпечує підвищення коефіцієнта ущільнення порівняно з неадаптивним методом ущільнення на основі лінійної форми Фібоначчі, який передбачає тільки одну модель джерела даних й одне правило кодування для кожного блоку вхідних даних.

Список літератури

1. Huffman D.A. A method for the construction of minimum redundancy codes / Huffman D.A. // Proceedings of the Institute of Electrical and Radio Engineers. – 1952. – V. 40, 9. – P. 1098-1101.

2. Cormack G.V. Data compression using dynamic Markov modeling / Cormack G.V., Horspool R.N. // Comput. J. – 1987. – V. 30, 6. – P. 541-550.

3. Storer J. A. Data Compression: Methods and Theory / Storer J. A. //Computes Science Press. - 1988. - V. 47, 1. – P. 23-29.

4. Ziv J. Compression of individual sequences via variable-rate coding / Ziv J., Lempel A. // IEEE Trans. Inf. Theory. – 1978. – V. IT-24, 5. – P. 530-536.

5. Welch T.A. A Technique for High Performance Data Compression / Welch T.A. // Computer/ - 1984. - №6. - P. 176-189.

6. Анисимов А. В. Обратное преобразование Фибоначчи / А. В. Анисимов, Я. П. Рындин, С. Е. Редько // Кибернетика. – 1982. - № 3. – С. 9-11.

7. Кшановський О.Д. Арифметичні методи ущільнення цифрової інформації / О.Д. Кшановський, С.В. Тітарчук, В.А. Лужецький //Вісник ВПІ. - 1999. - №5. - С. 83-87.

8. Кричевский Р.Е. Сжатие и поиск информации / Р.Е. Кричевский - М.: Радио и Связь, 1989. - 176 с.

9. Рябко Б.Я. Эффективный метод адаптивного арифметического кодирования для источников с большими алфавітами / Б.Я. Рябко, А.Н. Фионова //Проблемы передачи информации. - 1999. - Том 35, вып. 4. - С.34-39.

10. Rissanen J.J. Universal modeling and coding / Rissanen J.J., Langdon G.G. // IEEE Trans. Inf. Theory. – 1981. – V. IT-27, 1. – P. 12-23.

11. Лужецький В. А. Узагальнена модель адаптивного ущільнення даних / В. А. Лужецький, Л. А. Савицька, Шахзада Ашрафул Хок //Вісник ВПІ. - 1999. - №5. - С. 83-87.

Стаття надійшла:26.06.2012.

Відомості про авторів

Лужецький Володимир Андрійович – завідувач кафедри захисту інформації, Вінницький національний технічний університет, Хмельницьке шосе, 95, м. Вінниця, 21021, тел. 59-83-86

Савицька Людмила Анатоліївна – асистент кафедри обчислювальної техніки, Вінницький національний технічний університет, Хмельницьке шосе, 95, м. Вінниця, 21021, тел. 59-83-79