УДК 519.72

І.О. Завадський, к. ф.-м. н., доцент

Префіксні (Δ,*k*)-коди

Київський національний університет імені Тараса Шевченка, 03680, м. Київ, пр-т. Глушкова 4д, e-mail: zava@ukr.net

Igor O. Zavadskyi, Associate Professor, Ph. D.

Prefix (Δ,*k*)-codes

Taras Shevchenko National University of Kyiv, 03680, Kyiv, Glushkova av., 4d

Описано новий клас універсальних префіксних кодів, відмінною рисою яких є можливість використання кількох роздільників кодових слів в одному коді. Доведено такі властивості кодів описаного класу, як повнота та універсальність.

Ключові слова: префіксні коди,(Δ,k)-коди, повнота коду, універсальні коди.

The new family of universal prefix codes is introduced. The distinctive feature of this family is the use of multiple codeword delimiters in the same code. Such properties of representatives of this code family as completeness and universality are proved. The delimiters are the bit groups of the form 01…10 containg m_1 or… or m_t ones and the code with such delimiters is denoted by $C_{m_1,\dots,m_t}$. The bit sequence of the from 1…10 containg m_1 or… or m_t ones is called delimiting sequence for the code $C_{m_1,\dots,m_t}$. This code contains all codewords that consist of only some delimiting sequence and also all codewords satisfying the following conditions: (i) codeword doesn't start with the delimiting sequence; (ii) codeword ends with the bit sequence 0D and (iii) codeword doesn't contain the bit sequence 0D anywhere except the end, where D is the delimiting sequence. The name "(Δ,k)-codes" is due to representation of codewords as sequences of (Δ,k)-groups of bits. Each group is constructed as a concatenation of codes of Δ- and k-values. The value of Δ is represented by one bit in fact, while the code of k-value has the form 1…10 containing zero or more ones.

Key words: compression code, (2,3)-representation, text compression, prefix code.

Статтю представив д.ф.-м.н., проф. Анісімов А.В.

Вступ

У роботі [1] було розглянуто префіксний код C_1, що базується на поданні цілих чисел у двобазисній системі з основним базисом 2 і додатковим базисом 3, яке було введено у [2]. Цей код можна досить ефективно застосовувати для стискання інформації, передусім природномовних текстів. Оскільки код C_1 – префіксний, то він також і однозначно декодовний. Однак цей код не є повним. Це означає, що множину кодових слів можна розширити так, що властивість однозначної декодованості не втратиться. Таке розширення приводить до коду C_2, описаного в [1].

Код C_2 з [1] можна описати як код із роздільником 0110. У цій статті буде розглянуто цілу родину подібних стискальних кодів, кожен із яких має кілька роздільників вигляду 01…10, що містять m_1 або m_2 або… або m_t одиниць. Їх ми називатимемо (Δ,*k*)-кодами, а представника такого сімейства кодів позначатимемо як $C_{m_1,\dots,m_t}$. Наприклад, $C_{2,3}$ – це код із роздільниками 0110 та 01110.

Означення (Δ,*k*)-кодів

Означення 1. Код $C_{m_1,\dots,m_t}$ містить усі слова вигляду 1…10 із m_1 або… або m_t одиницями, а також усі слова, що задовольняють таким умовам:

- слово не починається з послідовності 1…10, що містить m_1 або m_2 або… або m_t одиниць;
- слово закінчується послідовністю 01…10, що містить m_1 або m_2 або… або m_t одиниць;
- слово не містить послідовності 01…10 із m_1 або m_2 або… або m_t одиницями ніде, крім кінця.

Тут і надалі ми припускаємо, що $0 < m_1 < \dots < m_t$.

Назва «(Δ,*k*)-коди» пов'язана з тим, що слова цих кодів можуть бути подані у вигляді послідовностей так званих *(Δ,k)-груп бітів*, що означуються в такий спосіб.

© І.О. Завадський, 2015

Означення 2. (Δ,k)-група утворюється з'єднанням двійкових кодів значень Δ і k. Для коду $C_{m_1,\dots,m_t}$ величина Δ може набувати 2^d+1 значень, де d – деякий параметр, $0 \le d < m_1$. Значення Δ подаються рядком "0" та всіма двійковими рядками довжини $d+1$ із першим символом "1". Параметр k – це будь-яке додатне ціле число. Його значення подається у вигляді послідовності 1…10, що містить k–1 одиниць.

Зв'язок між (Δ,k)-групами та кодами $C_{m_1,\dots,m_t}$ встановлює така теорема.

Теорема 1. Будь-яке слово з будь-якого коду $C_{m_1,\dots,m_t}$ може бути інтерпретоване як послідовність (Δ,k)-груп бітів.

Доведення. Якщо слово має вигляд 1…10, то, згідно з означенням 1, його довжина є не меншою за m_1+1. Але, згідно з означенням 2, довжина коду Δ не перевищує m_1 і код Δ не включає в себе всього слова вигляду 1…10, а отже, решта цього слова може бути інтерпретована як код k.

Розглядаючи інші кодові слова, зауважимо, що початок будь-якої двійкової послідовності може бути інтерпретовано і як код деякого значення Δ, і як код деякого значення k. Вилучимо з початку кодового слова код Δ, потім – код k і отримаємо коротше кодове слово, до якого застосуємо ту саму процедуру, аж поки не поділимо все слово на (Δ,k)-групи бітів або не отримаємо залишок, у який не вкладається ціла (Δ,k)-група. Подивимось, як згідно з цією процедурою оброблятиметься закінчення кодового слова, що має вигляд 01…10 і містить не менше m_1 одиниць. Перший біт "0" цього закінчення може бути або кінцем деякого коду k, або належати коду Δ. У першому випадку на останній ітерації отримуємо залишок 1…10 із не менш ніж m_1 одиницями, що, як показано вище, є повною (Δ,k)-групою. У другому випадку зазначимо, що код Δ містить не більш ніж m_1 бітів і після його вилучення отримаємо послідовність вигляду 1…10, яка представляє певне значення k. Таким чином, ситуація, коли на останній ітерації отримується залишок, у який не вкладається ціла (Δ,k)-група, неможлива, і будь-яке слово з будь-якого коду $C_{m_1,\dots,m_t}$ складається з цілої кількості (Δ,k)-груп бітів. ■

Зауважимо, що теорема 1 справедлива для будь-яких значень параметра d, що задовольняють нерівності $0 \le d < m_1$. Тому надалі для спрощення вважатимемо, що d=0, тобто код величини Δ складається з одного біта.

Алгоритми кодування та декодування

На вхід кодувального алгоритму надходить натуральне число x, двійковий вигляд якого – $x_n x_{n-1}\dots x_0$, а на виході отримуємо слово з коду $C_{m_1,\dots,m_t}$. Нехай J={j_1, j_2, …} – така нескінченна зростаюча послідовність натуральних чисел, що j_1 – найменше число, яке не належить множині {m_1,…,m_t}, а j_i – таке найменше число, що $j_i > j_{i-1}$ і $j_i \notin \{m_1,\dots,m_t\}$. Стандартний алгоритм кодування є таким:

1) $x \leftarrow x - 2^n$, тобто видаляємо найстарший біт числа x, який завжди дорівнює 1.

2) Якщо x=0, до рядка $x_{n-1}\dots x_0$, який містить лише нулі або порожній, дописуємо справа послідовність 1…10 з m_1 одиницями. Шукане кодове слово має вигляд $x_{n-1}\dots x_0$1…10. Кінець алгоритму.

3) Якщо двійкове подання x набуває вигляду рядка 1…10 або 0…01…10 із m_2 або… або m_t одиницями, це і є шукане кодове слово. Кінець алгоритму.

4) Розглядаємо кожну бітову послідовність вигляду 01…10, що не є розташованою в кінці слова послідовністю із m_2 або … або m_t одиницями. Якщо в цій послідовності d одиниць, замінюємо її подібною послідовністю 01…10, що містить j_d одиниць.

5) Якщо слово починається з послідовності 1…10, закінчується послідовністю 01…1, або цілком складається з послідовності вигляду 1…1, замінюємо кількість одиниць у цих послідовностях із d на j_d.

6) Якщо слово закінчується послідовністю 01…10, яка містить m_2 або … або m_t одиниць, це шукане слово. Кінець алгоритму.

7) Дописуємо до слова справа рядок 01…10, що містить m_1 одиниць.

Згідно з цим алгоритмом, якщо $x - 2^n \ne 0$, роздільник 01…10 із m_1 одиницями дописується штучно і тому має бути видалений під час декодування, а роздільники вигляду 01…10 із m_2 або … або m_t одиницями є інформативними частинами кодових слів і тому під час декодування повинні оброблятися. Якщо ж $x - 2^n = 0$, останні m_1+1 бітів вигляду 1…10 мають бути видалені.

Далі описано стандартний алгоритм декодування, на вхід якого подається кодове слово, а на виході отримуємо число.

1) Якщо кодове число має вигляд 0…01…10 або 1…10 і містить m_1 одиниць, видаляємо останні m_1+1 бітів і переходимо на крок 5.

2) Якщо число закінчується послідовністю 01…10, яка містить m_1 одиниць, видаляємо останні m_1+2 бітів.

3) Замінюємо кожну послідовність вигляду 01…10 із $j_d \in J$ одиницями, кінець якої не збігається з кінцем кодового слова, послідовністю подібного вигляду 01…10, що містить d одиниць.

4) Якщо слово починається з рядка 1…10 із $j_d \in J$ одиницями, замінюємо його подібним рядком 1…10, що містить d одиниць.

5) Дописуємо символ "1" до початку числа.

Повнота (Δ,*k*)-кодів

Із самого факту існування наведеної вище декодувальної процедури випливає, що (Δ,*k*)-коди є роздільними, однак це можна також довести, перевіривши виконання нерівності Крафта-Макмілана: $\sum_{c\in C} 2^{-|c|} \le 1$, де C – певний код. Крім того, усі (Δ,*k*)-коди мають ще одну важливу властивість: вони повні. Це означає, що до будь-якого коду з цього сімейства неможливо додати жодного слова так, щоб властивість роздільності не втратилася. Усі ці властивості доведено в теоремі 2.

Теорема 2. Будь-який (Δ,*k*)-код $C_{m_1,\dots,m_t}$ є роздільним і повним.

Доведення. Необхідною та достатньою умовою роздільності та повноти деякого коду C є виконання нерівності Крафта-Макмілана як рівності: $\sum_{c\in C} 2^{-|c|} = 1$. Якщо f_n – кількість кодових слів довжини n, то цю рівність можна переписати у вигляді

$$\sum_{n=1}^{\infty} 2^{-n} f_n = 1 \qquad (1)$$

Розглянемо (Δ,*k*)-код $C_{m_1,\dots,m_t}$. Для будь-якого $n \ge 2$ існує дві (Δ,*k*)-групи довжини n: 1…10 і 01…10, що містять n–1 та n–2 одиниць відповідно. Серед усіх (Δ,*k*)-груп ті, що містять m_1, m_2 або … або m_t одиниць, є термінальними, тобто такими, що можуть трапитися лише в кінці кодового слова. Отже, для коду $C_{m_1,\dots,m_t}$ існує 2t термінальних груп, довжини яких – m_1+1, m_1+2,…, m_t+1, m_t+2. Позначимо кількість термінальних груп довжини n через T_n. Очевидно, T_n дорівнює кількості входжень числа n в набір {m_1+1, m_1+2,…, m_t+1, m_t+2} (ця кількість становить 0, 1 або 2), а кількість нетермінальних груп довжини n дорівнює 2–T_n.

Розглянемо кодові слова довжини n, що містять принаймні дві (Δ,*k*)-групи. Кожне таке слово можна отримати додаванням зліва його першої нетермінальної (Δ,*k*)-групи до коротшого кодового слова. З іншого боку, додавання до будь-якого кодового слова зліва довільної нетермінальної групи утворює довше кодове слово. Якщо ж кодове слово містить тільки одну групу, то ця група є термінальною. Отже, врахувавши, що довжина найкоротшої (Δ,*k*)-групи становить 2, отримуємо таку рекурентну формулу для обчислення кількості кодових слів довжини n:

$$\begin{aligned} f_n &= T_n + \sum_{k=0}^{n-2}(2 - T_k) f_k = \\ &= T_n + 2(f_{n-2} + f_{n-3} + \cdots) - \\ &f_{n-(m_1+1)} - \cdots - f_{n-(m_t+1)} - \\ &f_{n-(m_1+2)} - \cdots - f_{n-(m_t+2)} \end{aligned} \qquad (2)$$

Застосуємо цю формулу для обчислення f_{n-1}:

$$\begin{aligned} f_{n-1} &= T_{n-1} + \sum_{k=0}^{n-3}(2 - T_k) f_k = \\ &= T_{n-1} + 2(f_{n-3} + f_{n-4} + \cdots) - \\ &f_{n-(m_1+2)} - \cdots - f_{n-(m_t+2)} - \\ &f_{n-(m_1+3)} - \cdots - f_{n-(m_t+3)} \end{aligned} \qquad (3)$$

Знайдемо праву частину (3) в (2) і замінимо її на f_{n-1}:

$$\begin{aligned} f_n &= T_n - T_{n-1} + 2f_{n-2} + f_{n-1} - \\ &f_{n-m_1-1} - \cdots - f_{n-m_t-1} + \\ &f_{n-m_1-3} + \cdots + f_{n-m_t-3} \end{aligned} \qquad (4)$$

Позначивши ліву частину (1) через s та врахувавши, що $f_0 = f_{-1} = \cdots = 0$, для будь-якого p>0 матимемо такі рівності: $\sum_{n=1}^{\infty} 2^{-n} f_{n-p} = 2^{-p} \sum_{n=1}^{\infty} 2^{-(n-p)} f_{n-p} = s2^{-p}$. Враховуючи їх та підставляючи вираз (4) в (1), отримаємо:

$$\begin{aligned} s &= \sum_{n=1}^{\infty} 2^{-n} f_n = \sum_{n=1}^{\infty} 2^{-n}(T_n - T_{n-1} + \\ &f_{n-1} + 2f_{n-2} - f_{n-(m_1+1)} - \cdots - f_{n-(m_t+1)} + \\ &f_{n-(m_1+3)} + \cdots + f_{n-(m_t+3)}) = \\ &\sum_{n=1}^{\infty} 2^{-n} T_n - \frac{1}{2}\sum_{n=1}^{\infty} 2^{-(n-1)} T_{n-1} + \\ &s(\frac{1}{2} + \frac{1}{2} - 2^{-m_1-1} - \cdots - 2^{-m_t-1} + \\ &2^{-m_1-3} + \cdots + 2^{-m_t-3}) \end{aligned} \qquad (5)$$

Взявши до уваги, що $2^{-m_i-3} - 2^{-m_i-1} = -3\cdot 2^{-m_i-3}$ для будь-якого i, $\sum_{n=1}^{\infty} 2^{-n} T_n = \sum_{n=1}^{\infty} 2^{-(n-1)} T_{n-1}$ і скоротивши s в обох частинах (5), отримаємо таку формулу:

$$3s \sum_{i=1}^{t} 2^{-m_i-3} = \frac{1}{2}\sum_{n=1}^{\infty} 2^{-n} T_n \qquad (6)$$

Оскільки довжини термінальних (Δ,*k*)-груп становлять $m_1 + 1, m_1 + 2, \dots, m_t + 1, m_t + 2$,

виконується рівність $\sum_{n=1}^{\infty} 2^{-n} T_n = \sum_{i=1}^{t}(2^{-m_i-1} + 2^{-m_i-2}) = \frac{3}{4}\sum_{i=1}^{t} 2^{-m_i}$. Тому нерівність (6) набуває вигляду $\frac{3}{8} s \sum_{i=1}^{t} 2^{-m_i} = \frac{3}{8}\sum_{i=1}^{t} 2^{-m_i}$, звідки випливає $s = 1$. ■

Універсальність (Δ,*k*)-кодів

Набір кодових слів із довжинами l_i ($l_1 \le l_2 \le \ldots$) називається універсальним у розумінні Еліаса [3], якщо існує така константа K, що для будь-якого скінченного розподілу ймовірностей $P=(p_1,\ldots,p_n)$, де $p_1 \ge p_2 \ge \ldots$ виконується нерівність

$$\sum_{i=1}^{n} l_i p_i \le K \cdot \max(1, E(P)), \qquad (7)$$

де $E(P) = -\sum_{i=1}^{n} p_i \log_2 p_i$ – ентропія розподілу P, а K – незалежна від P константа.

Теорема 3. Будь-який (Δ,*k*)-код $C_{m_1,\ldots,m_t}$ є універсальним.

Доведення. Оцінимо величину s_n – кількість кодових слів, довжина яких не перевищує n. Для цього скористаємося рівністю (4):

$$s_n = \sum_{i=1}^{n} f_i = \sum_{i=1}^{n}(T_i - T_{i-1}) + 2\sum_{i=1}^{n} f_{i-2} + f_{n-1} + \sum_{i=1}^{n-1} f_{i-1} - f_{n-m_1-1} - \cdots - f_{n-m_t-1} - -f_{n-m_1-2} - \cdots - f_{n-m_t-2} - \sum_{i=1}^{n-2}(f_{i-m_1-1} + \cdots + f_{i-m_t-1}) + \sum_{i=1}^{n}(f_{i-m_1-3} + \cdots + f_{i-m_t-3}) \qquad (8)$$

Зауважимо, що $\sum_{i=1}^{n}(T_i - T_{i-1}) = T_n - T_0 = T_n$, оскільки $T_0 = 0$; $2\sum_{i=1}^{n} f_{i-2} + \sum_{i=1}^{n-1} f_{i-1} = 3\sum_{i=1}^{n-2} f_i + 3f_0 + 2f_{-1} = 3\sum_{i=1}^{n-2} f_i$, оскільки $f_0 = f_{-1} = 0$; $-\sum_{i=1}^{n-2}(f_{i-m_1-1} + \cdots + f_{i-m_t-1}) + \sum_{i=1}^{n}(f_{i-m_1-3} + \cdots + f_{i-m_t-3}) = f_{-m_1-1} + \cdots + f_{-m_t-1} + f_{-m_1-2} + \cdots + f_{-m_t-2} = 0$, оскільки $f_i=0$, якщо $i < 0$. Таким чином, рівність (8) можна перетворити так:

$$s_n = T_n + 3\sum_{i=1}^{n-2} f_i - f_{n-m_1-1} - \cdots - f_{n-m_t-1} - f_{n-m_1-2} - \cdots - f_{n-m_t-2} > 3s_{n-m_t-2}.$$

Безпосередньо з означення 1 випливає, що $s_i > 0$, якщо $i \ge m_1$. Тому як тільки

$$n - m_t - 2 \ge m_1, \qquad (9)$$

виконується нерівність $s_n > \left(\sqrt[m_t+2]{3}\right)^{n-m_1}$. Позначивши $\sqrt[m_t+2]{3}$ через α, а $1/\alpha^{m_1}$ – через β, отримаємо нерівність $s_n > \beta\alpha^n$, де α>1 і β>0 – деякі константи, що залежать від m_1 і m_t. Таким чином, якщо l_i – це довжина i-го кодового слова у впорядкованому за зростанням довжин наборі кодових слів, то $i > s_{l_i-1} > \beta\alpha^{l_i-1}$. Звідси отримуємо

$$l_i < \log_\alpha(i/\beta) + 1 = \log_\alpha 2 \cdot \log_2 i - \log_\alpha \beta + 1 \qquad (10)$$

Ця нерівність виконується, якщо $l_i \ge m_1 + m_t + 2$ (умова (9)). Тоді $i \ge 2$, а отже, $\log_2 i \ge 1$ і нерівність (10) можна переписати у вигляді $l_i < (\log_\alpha 2 - \log_\alpha \beta + 1)\log_2 i = (\log_\alpha (2/\beta) + 1) \cdot \log_2 i$. Але оскільки величини p_i розташовано у незростаючому порядку, а їх сума дорівнює 1, то $p_i \le 1/i$, а отже, $\log_2 i \le -\log_2 p_i$; $l_i p_i < -(\log_\alpha(2/\beta) + 1)p_i \log_2 p_i$, а отже, константою K в нерівності (7) може бути величина $K_0 = \log_\alpha (2/\beta) + 1 = \log_\alpha 2\alpha^{m_1} + 1$. Враховуючи, що $\alpha = \sqrt[m_t+2]{3}$, маємо $K_0 = \log_3 2\,(m_t + 2) + m_1 + 1$. Якщо ж умова (9) не виконується, то ентропія $E(P)$ з нерівності (7) обмежена величиною $\log_2(m_1+m_t+2)$, а тому нерівність (7) виконується для певної величини $K=K_1$, що залежить від параметрів коду m_1 і m_t і не залежить від розподілу P. Таким чином, константою K з нерівності (7) може бути більша з величин K_0, K_1.■

Список літератури

1. *Завадський І.О.* Префіксний стискальний код на основі нижнього (2,3)-подання чисел / І.О. Завадський // Вісник Київського національного університету ім. Т. Шевченка. Серія: фіз.-мат. науки. – 2015. – №1. – с. 124-129.

2. *Анисимов А.В.* Помехоустойчивое префиксное кодирование на основе нижнего (2,3)-представления чисел / А.В. Анисимов, И.А. Завадский // Кибернетика и системный анализ. – 2014. – № 2. – С. 3–14.

3. *Elias P.* Universal codeword sets and representation of the integers / P.Elias // IEEE Trans. Inform. Theory – 1975. – v. 12, p. 194–203.

References

1. ZAVADSKYI, I.O. (2015) Prefix Compressing Code by Means of the 2,3-Representation of Numbers, *Bulletin of Taras Shevchenko National University of Kyiv. Series Physics & Mathematics*. No 3. pp. 124-129.

2. ANISIMOV, A.V., ZAVADSKYI, I.O. (2014) Pomekhoustojchivoe prefiksnoe kodirovanie na osnove nizhnego (2,3)-predstavlenija chisel. *Kibernetika i systemnyj analiz*. No 2, pp. 3–14.

3. ELIAS P. (1975) Universal codeword sets and representation of the integers. *IEEE Trans. Inform. Theory*. vol. 12, pp. 194–203.

Надійшла до редколегії 10.03.2015