

Д.т.н., професор Дичка І.А., магістрант Лавренчук С.С.

Національний технічний університет України
«Київський політехнічний інститут»

ВИЗНАЧЕННЯ ПОТУЖНОСТІ АЛФАВІТУ ДЛЯ УЩІЛЬНЕНОГО ПОДАННЯ АЛФАВІТНО-ЦИФРОВИХ ДАНИХ ПРИ ВИКОРИСТАННІ ТРИКОЛІРНОГО ГРАФІЧНОГО КОДУ

Abstract

*Ivan A. Dychka, prof., D.Sc.; Sergey Lavrenchuk, student
Determination alphabet power for sealing represents an alphanumeric data
using three-colours graphic code*

This paper concerns the task of determination alphabet power for sealing represents an alphanumeric data using three-colours graphic code. Compressing data before their submission in view of the bar code is mandatory steps to minimize their size. Choosing the optimal source alphabet allows the use of natural potential for data compression.

Вступ

Графічні коди (часто їх називають штрих-кодами), як лінійні так і двовимірні, набули значного розповсюдження у сучасному житті. Зараз важко уявити організацію складського обліку без використання комп'ютерного обладнання зі зчитувачами графічно-кодових позначок (ГК-позначок) та штрих-кодів. Але інколи виникає необхідність нанесення позначок на маленькі предмети – такі як корпуси мікросхем, медичні інструменти, невеличкі деталі. Тому постає задача мініатюризації ГК-позначок, які б забезпечували прийнятний рівень інформаційної ємності.

Більшість існуючих графічних кодів є двоколірними – зазвичай чорно-білими. Проте традиційне друкарське обладнання дозволяє отримувати проміжний колір – сірий, який, без особливих технічних труднощів, може ідентифікуватися зчитувачами ГК-кодів (сканерами). Використання третього кольору дозволяє збільшити інформаційну ємність ГК-позначок без зміни їх геометричних розмірів. Тому актуальним є дослідження застосування триколірних ГК-кодів в системах ідентифікації.

Постановка задачі

У сучасних системах автоматичної ідентифікації на основі графічного кодування даних вимагається, щоб графічно-кодований знак (ГК-знак) подавав від кількох сотень до кількох тисяч алфавітно-цифрових символів, тобто був не лише ідентифікатором об'єкта, а слугував би своєрідним переносним файлом даних (portable data file). Досягти цього можна лише забезпечивши високу інформаційну щільність даних, що підлягають поданню у графічно-кодованому вигляді.

Задача полягає у визначенні оптимальних потужностей вихідних алфавітів повідомлень, що можуть бути закодовані з максимально можливим ущільненням даних у вигляді двовимірної ГК-позначки використовуючи три кольори при кодуванні даних.

Методика проведення дослідження

Крім використання трьох кольорів при поданні даних у вигляді графічних кодів необхідно шукати додаткові можливості для ущільнення даних (структурні способи).

При формуванні графічно-кодованого повідомлення можуть подаватися будь-які алфавітно-цифрові дані, що представлені розширеним ASCII (комп'ютерним алфавітом). Але у більшості випадків під час створення деякої множини алфавітно-цифрових повідомлень може використовуватися обмежений набір символів з розширеного ASCII, який визначимо як вихідний алфавіт повідомлень.

Наприклад, при представленні деякої групи об'єктів можуть використовуватися лише великі літери латинської абетки (26 символів), 10 цифрових символів (0-9) та чотири розділові знаки (. , : -). Потужність такого алфавіту становить $26+10+4=40$ символів.

Таким чином якщо потужність алфавіту повідомлень менша за 256 (потужність розширеного ASCII) та відмінна від степеня трійки, то з'являється можливість закодувати повідомлення з ущільненням даних.

Під ущільненням даних будемо розуміти таке представлення алфавітно-цифрових повідомлень (тексту), при якому для повідомлення завдовжки k алфавітно-цифрових символів, що використовує алфавіт A потужності P_A , довжина $D(k)$ результуючої трійкової послідовності, отриманої в наслідок деякого перетворення вихідного тексту, буде задовольняти умову

$$D(k) < k \log_3 P_A. \quad (1)$$

Для неущільнених повідомлень $D(k) = k \log_3 P_A$.

Ущільненому повідомленню буде відповідати ГК-позначка меншої площі, що може забезпечити можливість для мініатюризації позначок за фіксованого обсягу даних або ж ущільнення забезпечить більшу інформаційну ємність позначки за її фіксованих геометричних розмірів.

Для забезпечення надійного зберігання графічно-кодованих даних використовують завадостійкий код Ріда-Соломона, який є багатозначним кодом, тобто оперує з недвійковими символами. Мінімальними інформаційними одиницями (розрядами) для коду Ріда-Соломона є q -значні ($q > 2$) символи – елементи поля Галуа $GF(q)$. При цьому можливі два випадки:

- 1) q є простим числом, тоді операції завадостійкого кодування/декодування виконують за модулем простого числа (за модулем q);
- 2) q є степенем простого числа ($q = p^s$, де $p \geq 3, s \geq 2$), тоді операції виконують за модулем незвідного многочлена степеня s .

При формуванні матриці ГК-позначки зручно користуватись ГК-знаками. Доцільно щоб ГК-знак складався з s комірок, і подавав один q -значний ($q = 3^s$) розряд кодового слова коду Ріда-Соломона. Тоді множина всіх можливих ГК-знаків утворить алфавіт Ω графічного коду потужності $P_{\Omega} = 3^s$, який називають символікою графічного коду; s -розрядний трійковий код, що відповідає ГК-знаку, називатимемо кодовектором.

Алфавіт Ω складається з інформаційних Ω_{inf} та службових Ω_{cl} ГК-знаків, $\Omega = \{\Omega_{inf}\} \cup \{\Omega_{cl}\}$. Інформаційні ГК-позначки Ω_{inf} використовуються для подання на носії вхідних алфавітно-цифрових послідовностей, службові ГК-знаки Ω_{cl} – для налаштування сканера, перемикання режимів кодування даних, тощо.

Процес ущільнення передбачає роботу лише з інформаційними кодовекторами, що утворюють алфавіт Ω_{inf} потужності $P_{\Omega_{inf}}$, $P_{\Omega_{inf}} = P_{\Omega} - P_{cl}$. Таким чином, кодування алфавітно-цифрової послідовності, що використовує алфавіт A потужності P_A , зводиться до перетворення символів алфавіту A в символи алфавіту Ω_{inf} потужності $P_{\Omega_{inf}} = P_{\Omega} - P_{cl}$.

У загальному випадку таке перетворення має вигляд:

$$n(P_A) \rightarrow m(P_{\Omega_{inf}}) \quad (2)$$

тобто n символів алфавіту A перетворюють в m символів алфавіту Ω_{inf} .

Реалізуємо перетворення (2) як переведення чисел з системи числення з основою P_A в систему числення з основою $P_{\Omega_{inf}}$. Зважаючи на те, що максимальне значення n -розрядного числа в системі числення з основою P_A дорівнює $P_A^n - 1$, а максимальне значення m -розрядного числа в системі числення з основою $P_{\Omega_{inf}}$ дорівнює $P_{\Omega_{inf}}^m - 1$, перетворення для випадку ущільнення зводиться до вигляду:

$$P_A^n - 1 \leq P_{\Omega_{\text{inf}}}^m - 1, \text{ або} \\ P_A^n \leq P_{\Omega_{\text{inf}}}^m. \quad (3)$$

Таким чином, для того щоб перетворення (2) забезпечувало ущільнення даних, необхідно і достатньо, щоб виконувалась умова:

$$\begin{cases} P_A^n \leq P_{\Omega_{\text{inf}}}^m, \\ \lfloor n \log_3 P_A \rfloor > \lfloor m \log_3 P_{\Omega_{\text{inf}}} \rfloor \end{cases} \quad (4)$$

де $\lfloor n \log_3 P_A \rfloor$ – довжина неущільненої n -розрядної послідовності в системі числення з основою P_A , а $\lfloor m \log_3 P_{\Omega_{\text{inf}}} \rfloor$ – довжина результуючого m -розрядного числа в системі числення з основою $P_{\Omega_{\text{inf}}}$.

Виконання умови (4) при фіксованому (обраному) $P_{\Omega_{\text{inf}}}$ зводиться до знаходження P_A – потужності алфавіту A , тобто до знаходження основи P_A відповідної системи числення, при якій буде забезпечуватися ущільнення.

Показник ущільнення алфавітно-цифрових даних будемо оцінювати коефіцієнтом ущільнення:

$$U_s^{(P_{\Omega_{\text{inf}}})}(P_A) = \frac{\lfloor n \log_3 P_A \rfloor}{\lfloor m \log_3 P_{\Omega_{\text{inf}}} \rfloor},$$

(5)

який показує, у скільки разів довжина трійкового рядка, що відповідає алфавітно-цифровій послідовності завдовжки n символів з алфавіту A ($\lfloor n \log_3 P_A \rfloor$), більша за довжину трійкового рядка ($\lfloor m \log_3 P_{\Omega_{\text{inf}}} \rfloor$) отриманого після виконання перетворення (2).

Нехай $s=5$. Тоді $P_{\Omega}=3^5=243$.

Зарезервуємо 13 службових символів, $P_{\text{сл}}=13$. Тоді кількість інформаційних слів становить $P_{\Omega_{\text{inf}}}=P_{\Omega}-13=230$. Отже формула (4) зводиться до вигляду:

$$\begin{cases} P_A^n \leq 230^m, \\ \lfloor n \log_3 P_A \rfloor > 5m. \end{cases}$$

Розв'язуючи цю систему нерівностей при невеликих $m=1,2,3,4,5$ – що впливає з практичних потреб кодування даних та проаналізувавши розв'язки можна дійти таких висновків: існують потужності вихідних алфавітів, що забезпечують прийнятні коефіцієнти ущільнення. Наприклад для алфавіту потужністю 37 символів (цього достатньо для представлення латинського алфавіту та десяткових цифр) отримуємо коефіцієнт 1,2, а для алфавіту десяткових цифр – коефіцієнт 1,4.

Висновки

Попередній аналіз вихідних даних для підбору оптимального значення вихідного алфавіту дозволяє досягти в деяких випадках значного ущільнення даних (до 40%). Тому цей етап повинен передувати створенню

будь-якого ГК, адже дозволяє використовувати природній потенціал для оптимізації.

Література

1. Two-Dimensional Codes. – Data Capture Institute, 1995. – 30 p.
2. Берлекэмп Э. Алгебраическая теория кодирования.–М.:Мир, 1971.–477с.
3. Теория кодирования / Т. Касами, Н. Токура, Е. Ивадари, Я. Инагаки. М.:Мир. –1978. – 576 с.