КІБЕРНЕТИКА

УДК 51.681.3

А.В. АНІСІМОВ
Київський національний університет імені Тараса Шевченка, Київ, Україна
e-mail: *a.v.anisimov@knu.ua.*

І.О. ЗАВАДСЬКИЙ
Київський національний університет імені Тараса Шевченка, Київ, Україна
e-mail: *ihorzavadskyi@knu.ua.*

Т.С. ЧУДАКОВ
Київський національний університет імені Тараса Шевченка, Київ, Україна
e-mail: *timofey.chudakov@gmail.com.*

СТИСНЕННЯ ПРИРОДНОМОВНИХ ТЕКСТІВ РЕВЕРСНИМИ МУЛЬТИРОЗДІЛЬНИКОВИМИ КОДАМИ

Анотація. У статті досліджено бінарні реверсні мультироздільникові (РМР) стискальні коди. РМР-коди мають низку корисних властивостей, як-от: однозначна декодовність, повнота, універсальність, синхронізовність, розпізнавання за допомогою скінченного автомата, а також можливість швидкого пошуку даних у закодованому файлі. Побудовано просте монотонне відображення з множини цілих невід'ємних чисел на множину кодових слів, а на його основі — швидкий побайтовий декодувальний алгоритм. Комп'ютерні експерименти демонструють, що РМР-код можна декодувати майже з тією самою швидкістю, що й код SCDC й у рази швидше, ніж код Фібоначчі. Якщо порівняти з відомими кодами подібного типу, РМР-коди демонструють кращий коефіцієнт стиснення природномовних текстів (більш ніж у 4 рази ближче до ентропійної межі, ніж SCDC). Також описано технологію передоброблення природномовних текстів, яка в поєднанні з кодуванням РМР-кодами підвищує ефективність потужних сучасних архіваторів.

Ключові слова: стиснення, архіватор, код, мультироздільниковий.

ВСТУП

Стиснення великих текстових корпусів є одним із ключових елементів сучасних інформаційно-пошукових систем. Методи стиснення тексту можна розділити на дві групи: методи, що використовують окремі символи як елементи алфавіту, і методи, що використовують слова як атомарні символи. Загалом методи другої групи мають кращі коефіцієнти стиснення, тому зосередимо увагу саме на них. Добре відомі класичні методи статистичного кодування можна застосувати до стиснення тексту на рівні слів та забезпечити коефіцієнти стиснення, близькі до теоретичної межі, визначеної ентропією Шеннона. Йдеться про арифметичне кодування, коди на основі асиметричних систем числення [1] і певною мірою коди Хаффмана [2]. Однак велике значення має не тільки ступінь стиснення, але й такі характеристики, як можливість пошуку у стиснутому потоці даних, висока швидкість декодування та обмеження можливого поширення помилок. Як відомо, зазначені коди недостатньо задовольняють перелічені вимоги.

Альтернативний підхід полягає у використанні кодів змінної довжини із суфіксами-роздільниками, відомих як патерн-коди. Роздільники — це спе-