УДК 004.056.5 : 519.728

В. А. ЛУЖЕЦЬКИЙ, Т. М. ЧЕБОРАКА
Вінницький національний технічний університет, м. Вінниця

МЕТОДИ УЩІЛЬНЕННЯ ДАНИХ НА ОСНОВІ ВІДКИДАННЯ ПОСЛІДОВНОСТЕЙ НУЛІВ ТА ОДИНИЦЬ

Анотація. Запропоновано методи ущільнення даних без втрат, що базуються на використанні методів відкидання послідовностей однакових символів у старших, молодших, внутрішніх та старших і молодших розрядах. Характерною особливістю цих методів є розбиття вхідної послідовності даних на блоки та їх подальша обробка на бітовому рівні, внаслідок чого зменшується їх залежність від типу оброблюваних даних. Розроблено програмний засіб для проведення експериментального дослідження запропонованих методів. Наведено результати дослідження ефективності ущільнення при використанні файлів такого типу: *.doc, *.txt, *.bmp, *.gif, *.jpg, *.au, *.mp3, *.exe, *.mdb.
Ключові слова: ущільнення даних, послідовності однакових символів.
Аннотация. Предложены методы сжатия данных без потерь, основаные на использовании методов отбрасывания последовательностей одинаковых символов в старших, младших, внутренних, старших и младших разрядах. Характерной особенностью этих методов является разбиение входной последовательности данных на блоки и их дальнейшая обработка на битовом уровне, в результате чего уменьшается их зависимость от типа обрабатываемых данных. Разработан программный продукт для проведения экспериментального исследования предложенных методов. Приводятся результаты исследования эффективности сжатия при использовании файлов следующего типа: *.doc, *.txt, *.bmp, *.gif, *.jpg, *.au, *.mp3, *.exe, *.mdb.
Ключевые слова: сжатие данных, последовательности одинаковых символов.
Abstract. The lossless methods of data compression, based on the truncation of the same characters sequences in low, high, internal and low and high order positions are proposed. A characteristic feature of these methods is a partitioning of input data sequence into blocks and further processing at bit level that leads to decreasing their dependence on the type of data being processed. A software tool for the experimental investigation of the proposed methods is developed. The research results of the effectiveness of data compression with use of the file types *.doc, *.txt, *.bmp, *.gif, *.jpg, *.au, *.mp3, *.exe, *.mdb are presented.
Keywords: data compression, sequences of the same characters.

Вступ

Обсяги інформації, яка зберігається і передається, з кожним роком значно зростають, що вимагає збільшення обсягів пам'яті і наявності високошвидкісних каналів передавання в сучасних інформаційних системах. Це призводить до постійного збільшення вартості таких систем. Тому економічно вигідним є застосування методів ущільнення для зменшення обсягів інформації, що зберігається і передається.

Існує декілька підходів до ущільнення інформації, що породжують цілу низку методів ущільнення [1, 2], для реалізації яких, у свою чергу, використовується величезна кількість алгоритмів. Одні з них мають науковий характер, а інші знайшли практичну реалізацію у вигляді архіваторів.

Актуальність

Теоретичні дослідження і практика застосування архіваторів показали, що не існує універсального методу ущільнення, що забезпечував би однаковий ступінь ущільнення для різних типів даних. Тому наукові дослідження спрямовані на створення ефективних методів ущільнення певних типів даних. Однак дані навіть одного типу, з погляду ущільнення, мають різні властивості і характеристики.

Мета

Метою роботи є підвищення значення коефіцієнта ущільнення даних без втрат шляхом створення методів ущільнення на основі відкидання послідовностей однакових символів та засобів, що їх реалізують.

Задачі

1. Розробити методи ущільнення даних без втрат на основі відкидання послідовностей нулів та одиниць.
2. Провести експериментальне дослідження ефективності ущільнення запропонованих методів на тестових файлах різного формату та обсягу.
3. Виконати порівняльний аналіз отриманих результатів.

Метод ущільнення даних на основі відкидання послідовності однакових символів у старших розрядах

Пропонується метод відкидання послідовностей нулів та одиниць у старших розрядах. Для спрощення викладання матеріалу у подальшому використовуватиметься скорочена назва даного методу – метод СТ.

Суть методу полягає в тому, що підраховується кількість q однакових символів, що містяться в старших розрядах і залежно від цієї кількості, початковий блок або перетворюється відкиданням послідовності цих однакових символів, або залишається незмінним.

Оскільки після відкидання q нулів (одиниць) наступним символом завжди буде протилежний символ, то цей символ також можна відкинути. Для того, щоб відновити символи, що відкидаються

необхідно додатково вказати кількість відкинутих символів q, у вигляді двійкового коду розрядності $\log_2 n$, де n – розрядність початкового блоку.

Якщо виконується умова $q < \log_2 n + 1$, то блок даних зберігається без змін, до якого дописується один додатковий символ p зі значенням 0, – ознакою збереження початкового блоку даних без змін наведено на рис. 1.

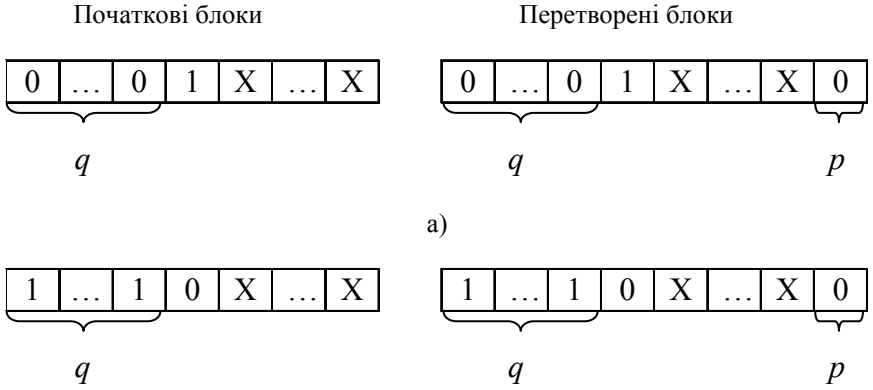

Рисунок 1 – Структури початкових та перетворених за методом СТ блоків (тип 0):
а) q нулів; б) q одиниць

При формуванні блоків типу 0 відбувається збільшення розрядності на один додатковий розряд порівняно з початковою структурою блоків.

При виконанні умови $q \geq \log_2 n + 1$, перетворений блок буде мати структуру, що наведена на рис. 2.

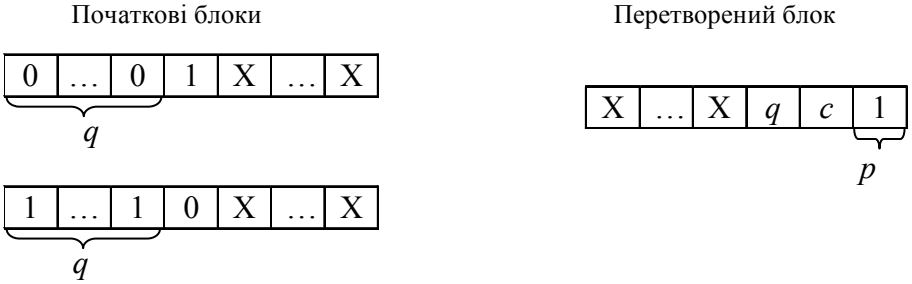

Рисунок 2 – Структури початкових та перетвореного за методом СТ блоку (тип 1)

Тут молодший розряд містить символ 1, що означає перетворення початкового блоку шляхом відкидання послідовності q однакових символів. Символ 2-го розряду c – вказує на тип символу, що складає послідовність, яка відкидається. Якщо $c = 1$, то це означає, що відкидається послідовність одиниць, а якщо $c = 0$, то – послідовність нулів. Поле q – двійковий код кількості однакових символів послідовності, що відкидається (розрядність коду $\log_2 n$). Поле $\mathrm{X...X}$ – код, що залишається без змін.

Метод ущільнення на основі відкидання послідовностей однакових символів у старших розрядах реалізується шляхом ініціалізації та створенням порожньої множини для значень ущільненої послідовності P.

Наступним кроком є зчитування блоків вхідної послідовності $D = \{d_1, d_2, ..., d_K\}$.

Далі для кожного блоку підраховується кількість однакових символів, що містяться в старших розрядах. При виконанні умови $q < \log_2 n + 1$ формуються блоки типу 0 за методом СТ ущільненої послідовності p_i^0. При виконанні умови $q \geq \log_2 n + 1$, формуються блоки типу 1 за методом СТ ущільненої послідовності p_i^1.

Метод ущільнення даних на основі відкидання послідовності однакових символів у молодших розрядах

Алгоритм ущільнення за методом відкидання послідовностей однакових символів у молодших розрядах (метод МР) є аналогічним методу СТ. Відмінність полягає лише у тому, що відбувається підрахунок кількості однакових символів у молодших розрядах блоку, що наведена на рис. 3.

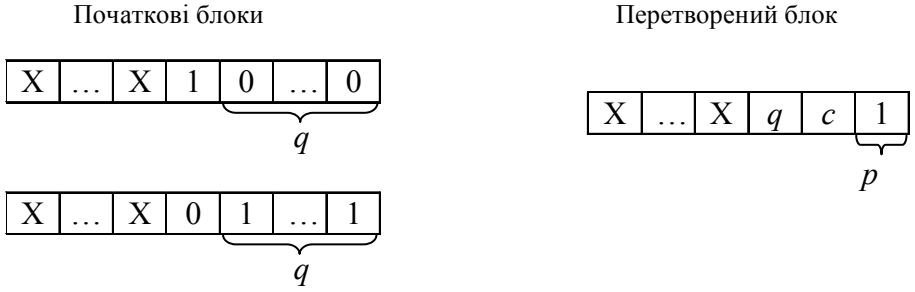

Рисунок 3 – Структури початкових та перетвореного за методом МР блоку (тип 1)

Метод ущільнення даних на основі відкидання послідовності однакових символів у внутрішніх розрядах

Пропонується метод відкидання послідовностей нулів та одиниць у внутрішніх розрядах. Для спрощення назви цього методу у подальшому використовуватиметься скорочена назва – метод ВР.

При виконанні умови $q < 2\log_2 n$ блок даних зберігається без змін, до якого дописується один додатковий символ p зі значенням 0, структури блоків наведені на рис. 4.

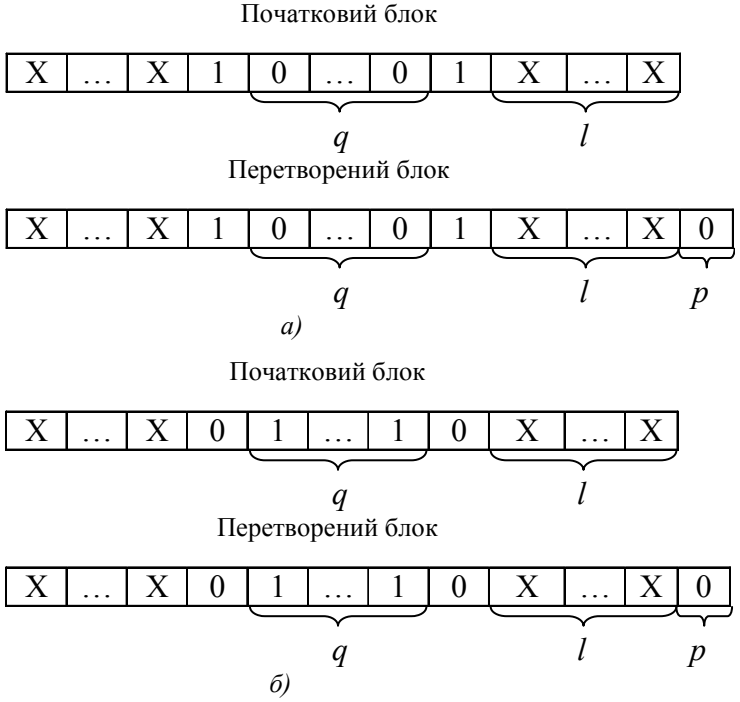

Рисунок 4 – Структури початкових та перетворених за методом ВР блоків (тип 0):
а) q нулів; б) q одиниць

При виконанні умови $q \geq 2\log_2 n$ перетворений блок матиме структуру, що наведена на рис. 5.

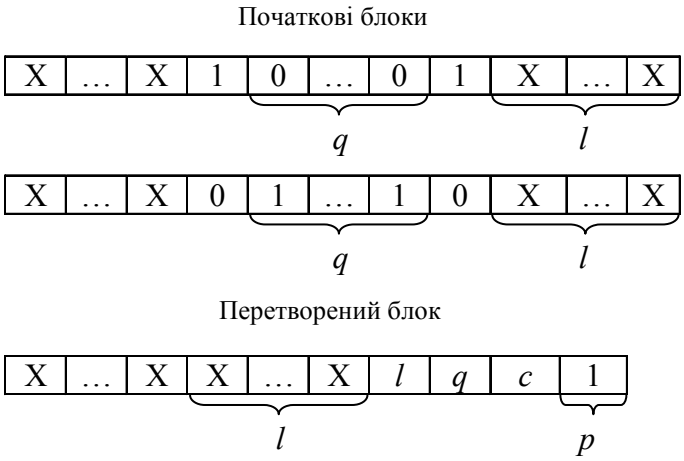

Рисунок 5 – Структури початкових та перетвореного за методом ВР блоків (тип 1)

Ця структура містить:

– $p = 1$ – ознаку перетворення;

– c – тип символу, що відкидається;

– поле q – двійковий код кількості однакових символів послідовності, що відкидається (розрядність коду $\log_2 n$);

– поле l – двійковий код кількості символів, що залишається без змін у молодших розрядах (розрядність коду $\log_2 n$);

– поле $\mathrm{X...X}$ – код, що залишається без змін.

Метод ущільнення даних на основі відкидання послідовності однакових символів у старших і молодших розрядах

Пропонується метод відкидання послідовностей нулів та одиниць у молодших і старших розрядах. Для спрощення викладання матеріалу у подальшому використовуватиметься скорочена назва цього методу – метод МСТ.

Суть методу полягає в тому, що підраховуються кількості q_l і q_h однакових символів, що містяться в молодших і старших розрядах відповідно, і залежно від цих кількостей початковий блок або перетворюється відкиданням послідовності однакових символів, або залишається незмінним.

Загальна структура початкового блоку, що підлягає ущільненню за методом МСТ наведена на рис. 6.

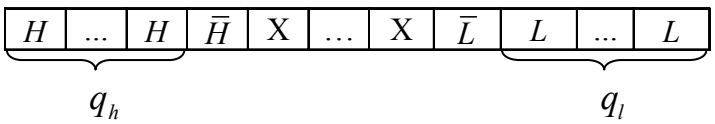

Рисунок 6 – Загальна структура початкового блоку, що підлягає ущільненню за методом МСТ

У наведеній структурі q_h – кількість однакових символів H, що відкидаються у старших розрядах, q_l – кількість однакових символів L, що відкидаються в молодших розрядах. Символи $\bar{H}$ та $\bar{L}$ протилежні символам H і L відповідно. Поле X…X – код, що залишається без змін.

Можливі варіанти структур початкового блоку наведено на рис. 7.

Для кожної структури, в свою чергу, можливі такі співвідношення кількостей однакових символів:

1) $q_h > q_l$;

2) $q_h < q_l$;

3) $q_h = q_l$.

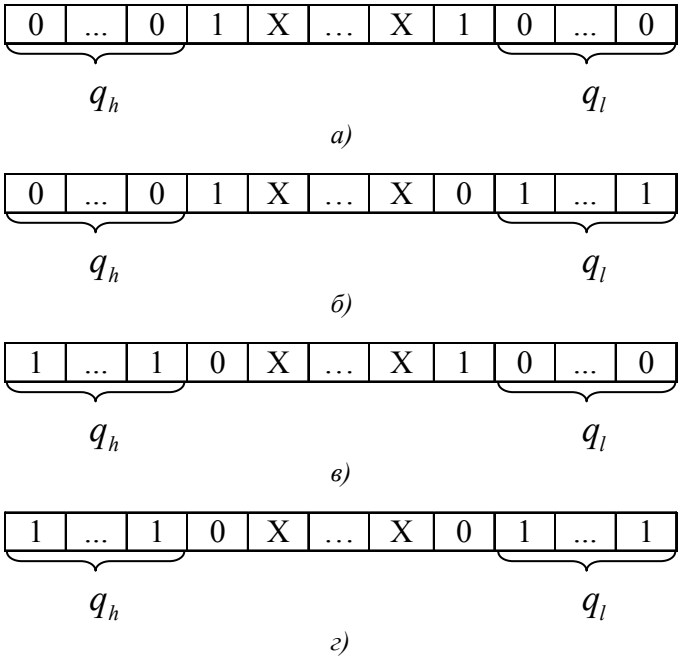

Рисунок 7 – Структури початкових блоків, що підлягають ущільненню за методом МСТ:
а) q_l і q_h нулів у молодших і старших розрядах відповідно; б) q_l одиниць у молодших і q_h нулів у старших розрядах; в) q_l нулів у молодших і q_h одиниць у старших розрядах;
г) q_l і q_h одиниць у молодших і старших розрядах відповідно

При виконанні умови $q_h + q_l \geq ((2\log_2 n) + 3)$ перетворений блок буде мати структури, що наведені на рис. 8.

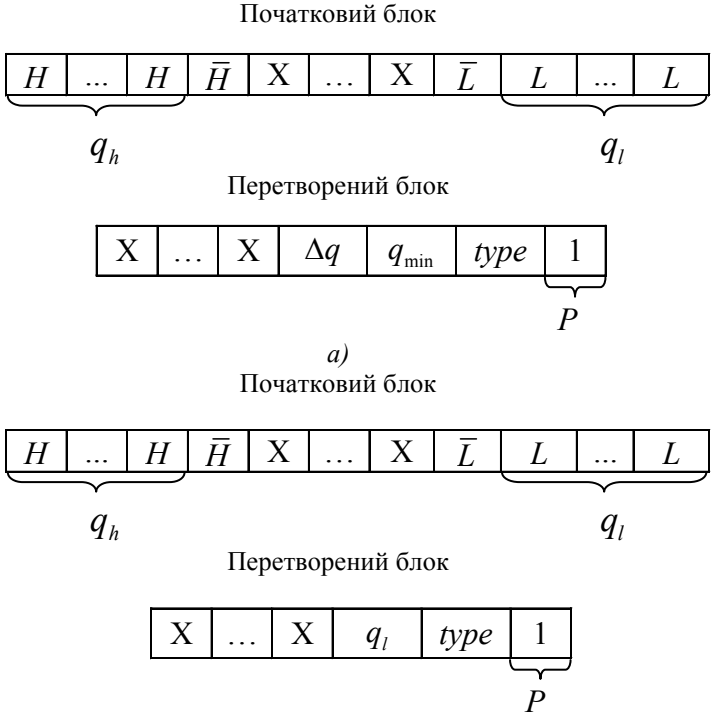

Рисунок 8 – Структури початкових та перетворених блоків, що підлягають ущільненню за методом МСТ:
а) $q_h > q_l$ або $(q_h < q_l)$; б) $q_h = q_l$

Ці структури перетворених блоків містять:

- $p = 1$ – ознаку перетворення;
- поле $type$ – 4-розрядний код типу перетворення;
- поле $q_{\min}$ розрядністю $\log_2 n$, в яке записується значення $\min(q_l, q_h)$;
- поле Δq розрядністю $\log_2 n$, в яке записується значення $|q_l - q_h|$;
- поле X...X – код, що залишається без змін.

У разі, коли виконуються співвідношення 1) і 2), структура перетвореного блоку має вигляд, наведений на рис. 8, а. Якщо $q_h = q_l$, то достатньо вказати одне із значень (нехай це буде q_l) і не потрібно вказувати Δq. З урахуванням цього маємо структуру перетвореного блоку, наведену на рис. 8, б.

Кількість можливих типів початкових структур, що підлягають перетворенню дорівнює 12. Коди типу структури $(type)$ і відповідні умови формування цих структур наведено в табл. 1.

Таблиця 1 – Позначення правил формування перетворених структур

H	L	Співвідношення q_h і q_l	Тип початкового блоку
0	0	$q_h > q_l$	0010
0	0	$q_h < q_l$	0001
0	0	$q_h = q_l$	0011
0	1	$q_h > q_l$	0110
0	1	$q_h < q_l$	0101
0	1	$q_h = q_l$	0111
1	0	$q_h > q_l$	1010
1	0	$q_h < q_l$	1001
1	0	$q_h = q_l$	1011
1	1	$q_h > q_l$	1110
1	1	$q_h < q_l$	1101
1	1	$q_h = q_l$	1111

Методика дослідження і програмний засіб для проведення досліджень

Вихідні дані, що підлягають ущільненню, будемо розглядається як послідовність символів 0 і 1. Ця послідовність розбивається на блоки, що містять однакову кількість символів, тобто виконується рівномірна розбивка [3]. Тоді структура вихідних даних матиме вигляд, наведений на рис. 9 [4, 5].

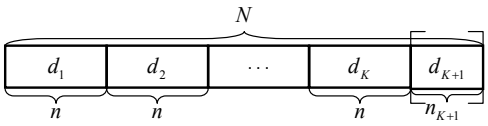

Рисунок 9 – Структура вихідних даних

Тут вихідна послідовність D із N символів розбита на блоки d_1, d_2, …, d_K $\left[, d_{K+1}\right]$ по n символів, де n кратне байту: $D = \left\{d_1, d_2, ..., d_K \left[, d_{k+1}\right]\right\}$.

Кількість блоків розрядністю n дорівнює $K = \lfloor N/n \rfloor$, де $\lfloor \ \rfloor$ означає округлення до меншого цілого. Якщо N ділиться на n точно, то всі блоки будуть мати однакову довжину n, і блок d_{K+1} буде відсутнім. Якщо результат ділення є нецілим числом, то блок d_{K+1} буде мати довжину $n_{K+1} = N \bmod n$.

Для того, щоб сформувати ущільнену послідовність *P,* необхідно поставити у відповідність кожному початковому блоку перетворений блок $d_i \rightarrow p_i$. Результатом перетворення вхідного потоку даних є ущільнена послідовність: $P = \left\{p_1, p_2, ..., p_K \left[, p_{K+1}\right]\right\}$.

Над блоком d_{K+1} перетворення не виконуються і він залишається без змін – $p_{K+1} = d_{K+1}$.

Для автоматизації експериментального дослідження запропонованих методів ущільнення на тестових файлах розроблено програмний засіб.

Під час проведення дослідження розрядність початкового блоку даних може приймати значення в діапазоні: 8, 16, 32, …, 8192. Дослідження відбувалося за такими показниками: коефіцієнт ущільнення k, тривалість ущільнення t_c, тривалість відновлення t_{dc}.

Результати дослідження запропонованих методів ущільнення за показниками k, t_c, t_{dc} виводяться у табличному та графічному вигляді (рис. 10).

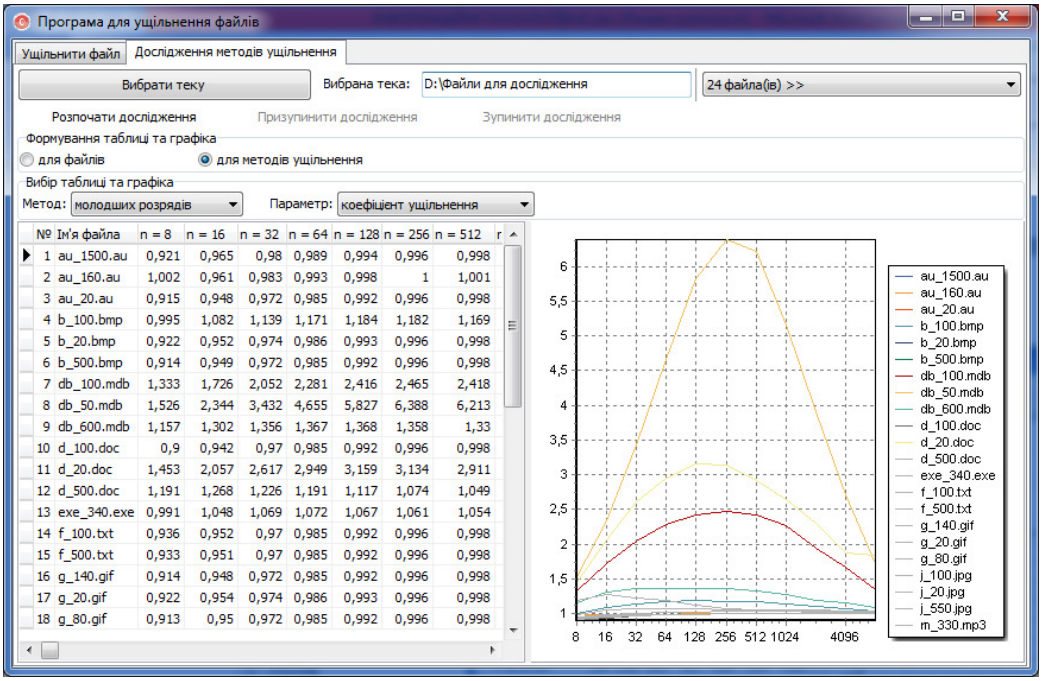

Рисунок 10 – Головне вікно програмного засобу

У процесі ущільнення вхідного файлу формується вихідний файл, що містить усю необхідну інформацію для відновлення початкового файлу.

Для проведення експериментальних досліджень запропонованих методів ущільнення було сформовано тестову вибірку із 24 файлів таких форматів та розмірів:

- текстового формату *.doc – 20 кБ, 100 кБ, 500 кБ;
- текстового формату *.txt – 100 кБ, 500 кБ;
- графічного формату *.bmp – 20 кБ, 100 кБ, 500 кБ;
- графічного формату *.gif – 20 кБ, 80 кБ, 140 кБ;
- графічного формату *.jpg – 20 кБ, 100 кБ, 550 кБ;
- музичного формату *.au – 20 кБ, 160 кБ, 1500 кБ;

– музичного формату *.mp3 – 40 кБ, 330 кБ, 500 кБ;
– формату виконуваних файлів *.exe – 340 кБ;
– формату бази даних *.mdb – 50 кБ, 100 кБ, 600 кБ.

Під час експериментальних досліджень вміст вихідних файлів розбивався на блоки розрядності $n = 8,16,...,2048$.

Результати досліджень коефіцієнта ущільнення наведено у табл. 2 та представлено у вигляді графіка на рис. 11. У таблиці для кожного значення коефіцієнта ущільнення також зазначено в дужках розрядність блоків, при якій було досягнуто цей коефіцієнт.

Таблиця 2 – Результати дослідження коефіцієнта ущільнення

Формат та обсяг файлу	Метод ущільнення			
	МР	СТ	ВР	МСТ
au, 160 кБ	1,002 (8)	1,187 (16)	1,012 (32)	1,007(16)
bmp, 100 кБ	1,184 (128)	1,181 (128)	1,174 (128)	1,198(256)
mdb, 100 кБ	2,465 (256)	2,484 (256)	2,485 (256)	2,651(1024)
mdb, 50 кБ	6,388 (256)	7,23 (512)	6,468 (256)	8,388(1024)
mdb, 600 кБ	1,368 (128)	1,374 (64)	1,414 (128)	1,386(256)
doc, 20 кБ	3,159 (128)	3,327 (128)	3,262 (256)	3,389(256)
doc, 500 кБ	1,268 (16)	1,43 (16)	1,25 (16)	1,196(16)
exe, 340 кБ	1,072 (64)	1,082 (32)	1,081 (512)	1,082(64)
mp3, 40 кБ	1,033 (1024)	1,031 (256)	1,037 (512)	1,034(2048)

З табл. 2 та рис. 11 видно, що для більшості файлів, що піддаються ущільненню запропонованими методами, найкращі результати демонструє метод ущільнення – МСТ. Лише для файлів (au, 160 кБ) та (doc, 500 кБ) метод СТ в незначній мірі дав кращі результати, ніж метод відкидання у старших і молодших розрядах: 1,187 проти 1,007 і 1,43 проти 1,196 відповідно.

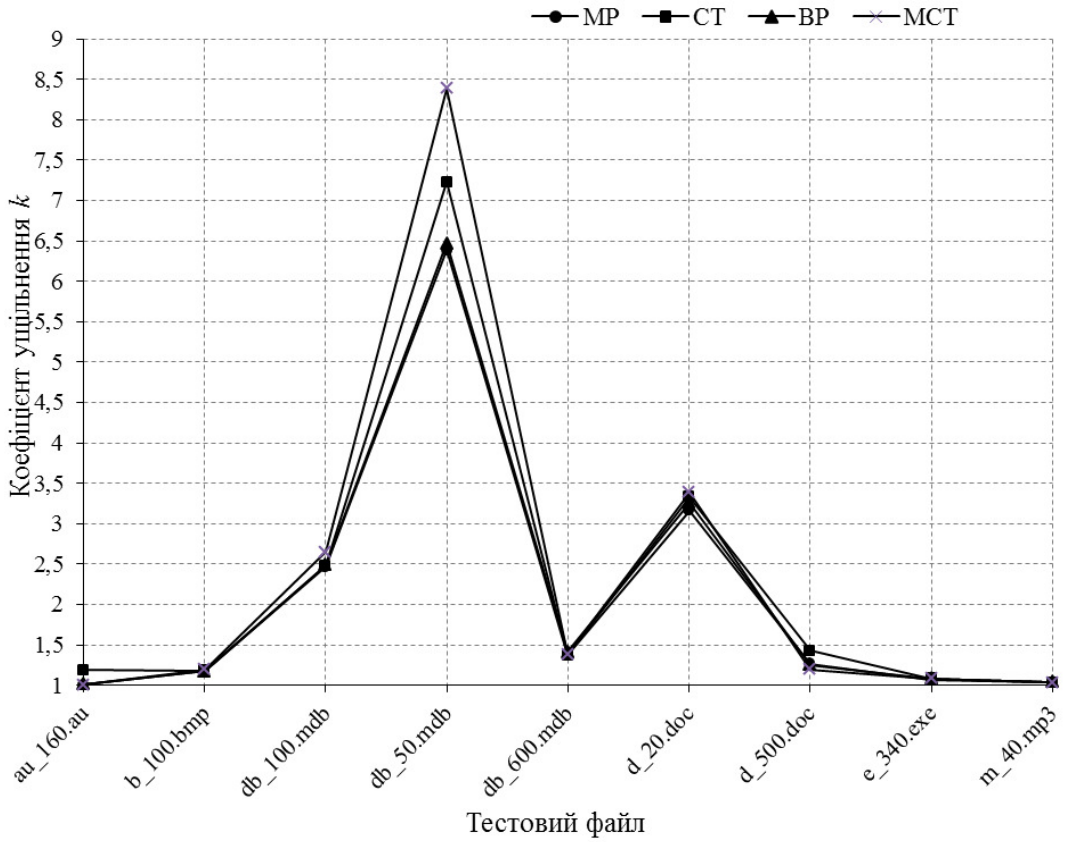

Рисунок 11 – Графіки результатів дослідження коефіцієнта ущільнення

Таким чином, проведені експериментальні дослідження запропонованих методів ущільнення показали, що ці методи дозволяють ущільнювати файли різних форматів та розмірів. В результаті досліджень найбільший коефіцієнт ущільнення із запропонованих методів продемонстрував метод відкидання послідовностей однакових символів у старших і молодших розрядах, який дав максимальний приріст значення коефіцієнта ущільнення у 16,01% порівняно із найкращими значеннями інших методів.

Щодо тривалості ущільнення та відновлення файлів запропонованими методами, то експериментальне дослідження на комп'ютері типу IBM/PC з процесором Intel Core i3 M370 (2.4GHz) і ОЗП розміром у 3 ГБ показало, що для найбільшого файлу (au, 1500 кБ) процедура ущільнення найдовше тривала при найменшій розрядності блоків: від 0,53 с. для методу МР та методу СТ до 0,5 с., – а процедура відновлення – 0,46 с. для методу ВР та 0,39 с. для методу МСТ.

Висновки

1. Запропоновано методи ущільнення даних на основі відкидання послідовностей нулів та одиниць у молодших, старших, внутрішніх та старших і молодших розрядах. Ці методи на відміну від існуючих обробляють вихідну послідовність даних на бітовому рівні, що дозволяє зменшити їх залежність від типу даних.
2. Проведено експериментальне дослідження коефіцієнта ущільнення запропонованих методів на тестових файлах різного формату та обсягу за допомогою розробленого програмного засобу. Це дослідження показало, що запропоновані методи ущільнення дозволяють ущільнювати файли різних форматів та розмірів.
3. Найбільший коефіцієнт ущільнення із запропонованих методів продемонстрував метод відкидання послідовностей однакових символів у старших і молодших розрядах, який дав максимальний приріст значення коефіцієнта ущільнення у 16,01% порівняно із найкращими значеннями інших методів.

Список використаних джерел

1. Ватолин Д. Методы сжатия данных. / Д. Ватолин, А. Ратушняк, М. Смирнов, В. Юкин – М.: ДИАЛОГ-МИФИ, 2002. – 384 с.
2. Salomon D. Handbook of Data Compression / D. Salomon, G. Motta. – London: Springer, 2010. – 1361 p.
3. Лужецький, В. А. Узагальнена модель адаптивного ущільнення даних / В. А. Лужецький, Л. А. Савицька, Ш. А. Хок // Інформаційні технології та комп'ютерна інженерія. – 2009. – № 1. – С. 56-63.
4. Лужецький В. А. Дослідження методу ущільнення даних на основі методу відкидання послідовностей нулів та одиниць / В. А. Лужецький , Т. М. Алєксєєва // Методи та засоби кодування, захисту й ущільнення інформації: міжнар. наук.-практ. конф., 20–22 квітня 2011 р.: тези доп., – Вінниця, 2011. – С.164-165.
5. Лужецький В. А. Дослідження числових моделей даних / В. А. Лужецький, А. В. Кульчицький , Т. М. Алєксєєва // Інформаційні технології та комп`ютерна інженерія. – 2010. – № 3. – С. 50–56.

Інформація про авторів

Лужецький Володимир Андрійович – д.т.н., професор, завідувач кафедри захисту інформації, Вінницький національний технічний університет, вул. Хмельницьке шосе 95, м. Вінниця.
Чеборака Тетяна Михайлівна – аспірант кафедри захисту інформації, Вінницький національний технічний університет, вул. Хмельницьке шосе 95, м. Вінниця.