ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ

УДК 681.3

В. А. ЛУЖЕЦЬКИЙ, А. В. КУЛЬЧИЦЬКИЙ, Т. М. АЛЄКСЄЄВА

Вінницький національний технічний університет, Вінниця

ДОСЛІДЖЕННЯ ЧИСЛОВИХ МОДЕЛЕЙ ДАНИХ

Анотація: Розглядаються числові моделі даних, що підлягають ущільненню, особливістю яких є те що будь-який тип даних розглядається як цілі додатні числа. Наводяться результати дослідження числових моделей, що відповідають файлам типу *.doc, *.exe, *.pdf, *.zip.

Ключові слова: Числові моделі даних.

Вступ

Кількість потрібної людині інформації неухильно зростає. Кількість пристроїв для зберігання даних і пропускна спроможність каналів зв'язку також зростають. Однак кількість інформації зростає швидше. У цієї проблеми є три рішення. Перше - обмеження кількості інформації. На жаль, воно не завжди прийнятно. Друге - збільшення кількості носіїв інформації та пропускної здатності каналів зв'язку. Це рішення пов'язане з матеріальними витратами, причому іноді досить значними. Третє рішення - використання ущільнення інформації. Це рішення дозволяє в кілька разів скоротити вимоги до кількості пристроїв зберігання даних і пропускної здатності каналів зв'язку без додаткових витрат (за винятком витрат на реалізацію алгоритмів стиснення) [1].

Саме завдяки необхідності використання ущільнення інформації методи ущільнення досить поширені. Однак існують дві серйозні проблеми. По-перше, широко застосовувані методи ущільнення, як правило, застаріли і не забезпечують достатнього зменшення надлишковості даних. У той же час вони вбудовані у велику кількість програмних продуктів і бібліотек і тому будуть використовуватися ще досить довгий час. Другою проблемою є часте застосування методів ущільнення, які не відповідають характеру даних. Вирішення цих проблем дозволяє різко підвищити ефективність застосування алгоритмів ущільнення. [2]

Тому виникає потреба у створені універсальних (адаптивних) методів ущільнення. Адаптивні методи ущільнення (adaptive encoding) - методи ущільнення даних, які заздалегідь не налаштовуються на певний тип даних [3]. Вони налаштовуються на будь-який тип даних, домагаючись максимального скорочення їх обсягу.

Одним з найбільш важливих досягнень у теорії ущільнення даних за останні десятиліття є розділення процесу ущільнення на дві частини: кодування, що перетворює дані в ущільнений потік бітів, і моделювання, яке підготовляє інформацію до нього [4].

Для того щоб досягти великої адаптивності ущільнення, потрібно на етапі моделювання джерела даних проводити дослідження, які дали б змогу оцінювали утворену модель даних і вибрати оптимальний метод кодування цих даних.

Метою роботи є підвищення ефективності ущільнення даних за рахунок використання числових моделей та їх дослідження перед ущільненням даних.

Для досягнення поставленої мети необхідно розв’язати такі задачі:

- розробити числові моделі;
- дослідити числові моделі.

Числові моделі даних

Моделювання джерела даних - процес створення моделі джерела даних, за допомогою застосування правил і певної техніки моделювання.

В обчислювальній техніці всі дані представляються у вигляді послідовності символів 0 і 1. Тому, для спрощення роботи, при побудові числових моделей вхідні дані також розглядатимуться як послідовності символів 0 і 1.

При моделюванні використовується рівномірне розбиття на блоки, тобто вхідна послідовність символів розбивається на блоки, що містять однакову кількість символів (рис. 1).

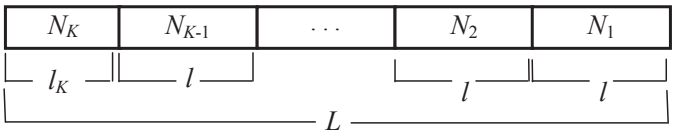


Рисунок 1 – Структура вхідних даних

Тут послідовність із L символів розбита на блоки по l символів. Кількість блоків дорівнює $K =]L/l[$, де $]\ [$ означає округлення до більшого цілого. Якщо L ділиться на l точно, то всі блоки будуть мати однакову довжину l. Коли результат ділення є нецілим числом, то один блок буде мати довжину $l_K = L - K \cdot l$.

$$M = \{a_0, a_1, \ldots, a_{L-1}\},$$

де a_i - символ алфавіту $\mathbf{A} = \{0,1\}$.

$$M = \{M_0^*, M_1^*, \ldots, M_{K-1}^*\},$$

де $M_j^* = \{a_{j\cdot l+0}, a_{j\cdot l+1}, a_{j\cdot l+2}, \ldots, a_{j\cdot l+l}\}$.

Далі кожному i-му блоку (i=1;2;..K) ставиться у відповідність число N_i.

$$N_i = f(M_i^*).$$

Функція $f()$ може бути описана формулою:

$$N_i = \sum_{j=0}^{l-1} a_{ij} w_j \qquad (1),$$

де a_{ij} - j-й символ i-го блоку;

w_j - числовий еквівалент j-ї позиції (розряду) коду.

Використавши функцію $f()$ для кожного блоку M_i^* ($i = 0,1, \ldots, K$), буде отримано послідовність з K чисел, що є числовою моделлю вхідних даних.

Із формули (1) випливають три основні підходи до побудови числових моделей, що адаптуються. Ці підходи базуються на:

1) варіюванні довжини блоків;
2) використанні системи числових еквівалентів;
3) перестановках символів блоку.

Перший підхід є найпростішим. Разом із зміною довжини блоків l змінюється значення чисел числової моделі, а також їх кількість.

У другому підході числова модель змінюється внаслідок використання різних систем числових еквівалентів. Забезпечити велику кількість можливих числових еквівалентів можна використовуючи формулу:

$$w_i = \sum_{j=0}^{i-1} w_j + m; \quad w_0 = 1; \; w_1 = k; \; m = 1,2,\ldots; \; k = 2,3,\ldots$$

Таким чином, є можливість формувати $m \cdot k$ систем числових еквівалентів для моделювання.

У третьому підході числова модель змінюється за рахунок перестановок символів у блоках, на які розбиті вхідні дані. Ці перестановки є оборотними, тобто перестановки після яких можна повністю відновити початкові дані. Прикладом таки перестановок є транспозиція та циклічний зсув.

Транспозиція – це перестановка в парах або в групах бітів, яка починається з заданого біту (рис. 2).

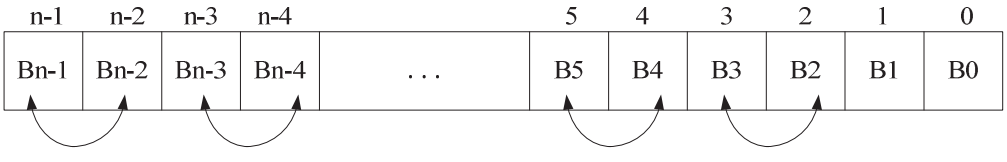


Рисунок 2 – Приклад транспозиції

Розглянуті вище підходи до побудови числових моделей роблять ці моделі адаптивними. Таким чином можна не тільки підбирати тип кодування під модель даних, але й модель даних підлаштовувати під певний тип кодування.

Для вибору найбільш ефективного методу кодування, потрібно дослідити властивості числових моделей вхідних даних.

Методика дослідження і програмний засіб для проведення досліджень

Оскільки числова модель – це послідовність чисел, то в основі дослідження лежить обчислення статистики появи чисел з діапазону $[0\ ;\ 2^l - 1]$, де l - кількість розрядів у блоці.

Нехай при моделюванні джерело даних розбивається на блоки з l біт - $\{a_1, a_2, a_3, ..., a_l\}$.

Далі кожному з блоків ставиться у відповідність число N_i, де $N_i \in [0; 2^l - 1]$.

Далі для кожного числа з діапазону $[0\ ;\ 2^l - 1]$ підраховується кількість його появ у числовій моделі. Результатом даного дослідження є графік, на горизонтальній вісі якого відкладаються числа від 0 до $2^l - 1$, а по вертикальній вісі кількість їх появи у числовій моделі.

Для автоматизації дослідження розроблено програмний засіб, що реалізує вище описаний підхід.

Даний програмний засіб виконує такі функції.

1. Моделювання вхідної інформації з використанням вище зазначених перетворень.
2. Проведення дослідження утвореної числової моделі.
3. Візуалізація результатів досліджень.

Даний програмний засіб перед моделюванням даних дозволяє задавати основні параметри моделювання (рис. 3), що дає змогу керувати числовими моделями, які утворюються.

До основних параметрів належать:

- розмір блоку;
- параметри траспозиції;
- параметри циклічного зсуву.

Після проведення дослідження результати виводяться у вигляді графіка (рис. 3).

Горизонтальна вісь (рис. 3, а) графіку з результатами представляє значення чисел з діапазону $[0\ ,\ 2^l - 1]$, де l - кількість розрядів у блоці. Оскільки діапазон дуже великий, то зображення кожного можливого значення на горизонтальній осі створить незручність перегляду результатів. Для вирішення даної проблеми діапазон розбивається на 600 піддіапазонів, розмір яких є різним для різних розмірів блоку (рис. 3, б).

Вертикальна вісь (рис. 3, в) ставить у відповідність кожному числу кількість його появ у числовій моделі. Оскільки числа розбиті на піддіапазони, то на вертикальній осі вказується частота появи чисел, що належать піддіапазону.

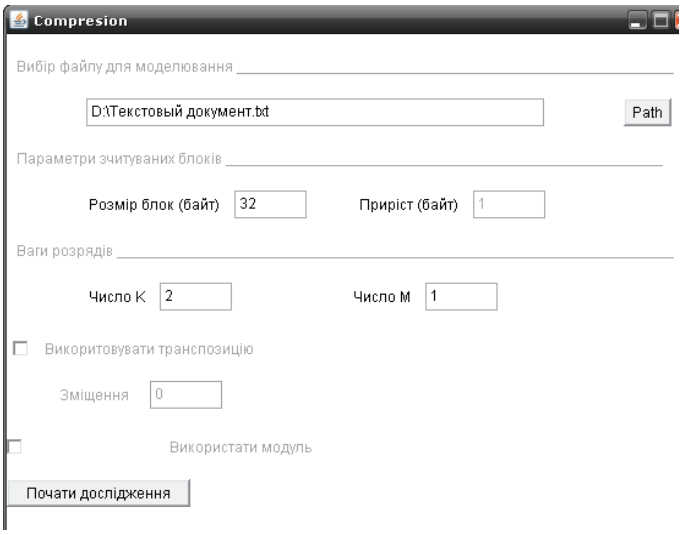


Рисунок 3 – Вигляд вікна з налаштуваннями моделювання

У верхній частині вікна вказуються параметри файлу, параметри моделювання та параметри перетворень (рис. 4, г).

Аналіз отриманих результатів

Під час досліджень числових моделей перевірялась залежність результатів від такого:

- розрядність блоків при моделюванні;
- тип даних;
- вміст файлу.

Результати дослідження впливу зміни розрядності блоків при моделюванні на числову модель представлені на рис.5, рис. 6 і рис. 7.

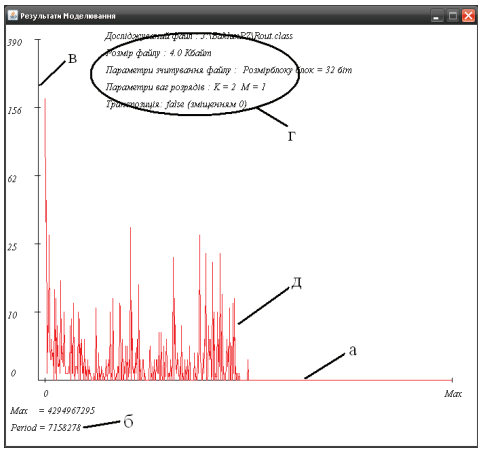
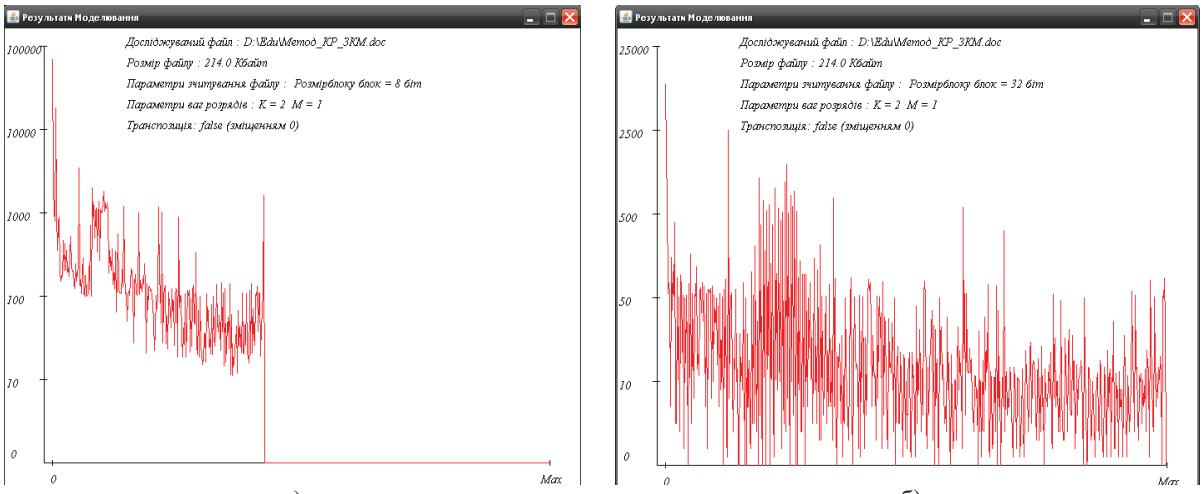


Рисунок 4 – Вигляд вікна з результатами досліджень:

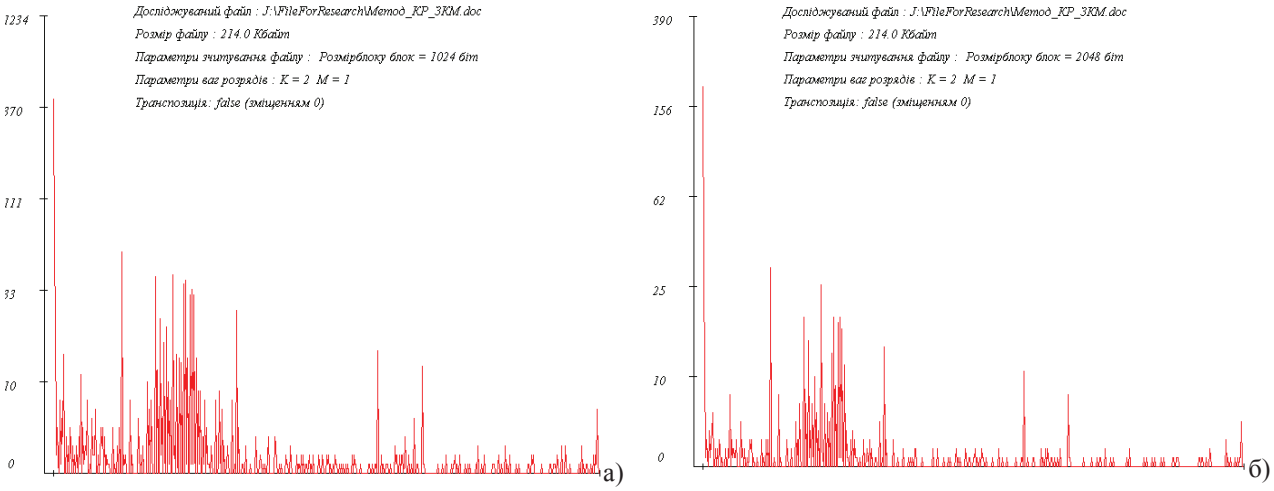
а) горизонтальна вісь; б) період; в) вертикальна вісь; г) параметри моделювання; д) результат дослідження



а) б)

Рисунок 5 – Результати досліджень для розрядності блоків:

а) 8 біт; б) 32 біт;



а) б)

Рисунок 6 – Результати досліджень для розрядності блоків:

а) 1024 біт; б) 2048 біт;

При зміні розрядності блоків числова модель радикально змінюється.

Для блоків малої розрядності спостерігається більш рівномірний розподіл по всіх піддіапазонах.

Збільшення розрядності блоків до 1024 і 2048 приводить до зменшення кількості значень у піддіапазонах, що наближаються до максимального значення, і появи екстремумів у піддіапазонах, що наближаються до нульового значення.

Подальше збільшення розрядності блоку впливає таким чином, що графіки розподілу мають екстремуми в певних піддіапазонах і не мають значень в інших.

Також було проведено дослідження впливу типу даних на числову модель. При цьому для однакових параметрів моделювання здійснювалась побудова числових моделей для файлів різних типів з приблизно однаковими розмірами.

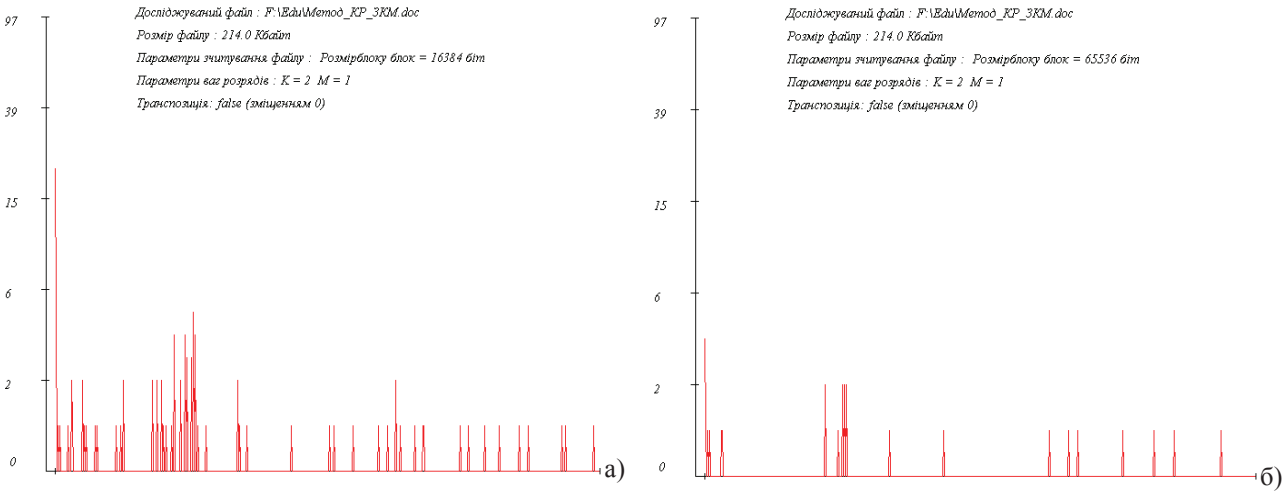


Рисунок 7 – Результати досліджень для розрядності блоків:

а) 16384; б) 65534 біт

Результати такого дослідження наведено на рис. 9 і 10.

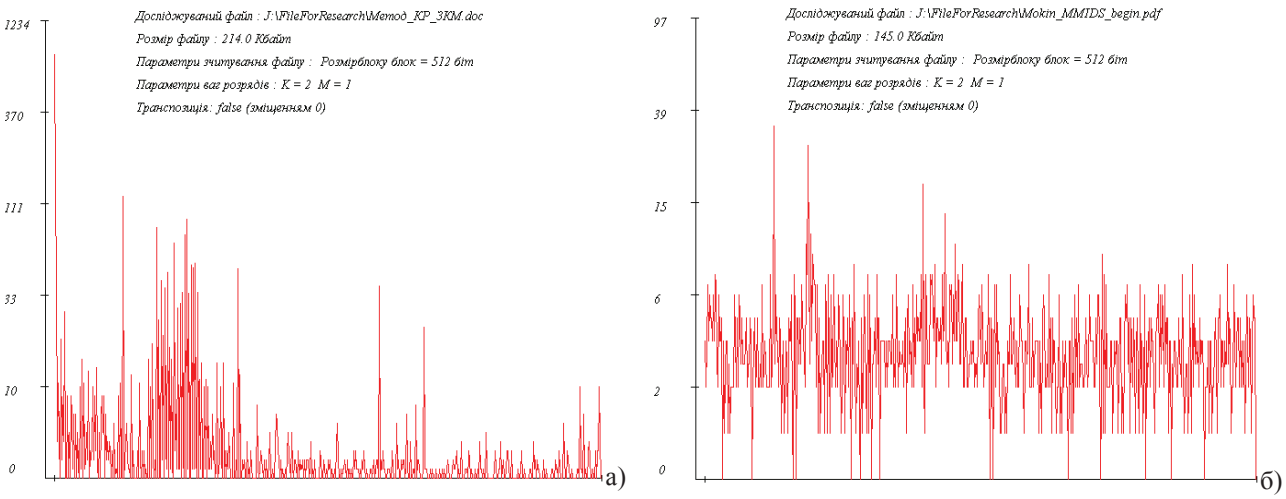


Рисунок 8 – Результати досліджень моделей для різних типів даних:

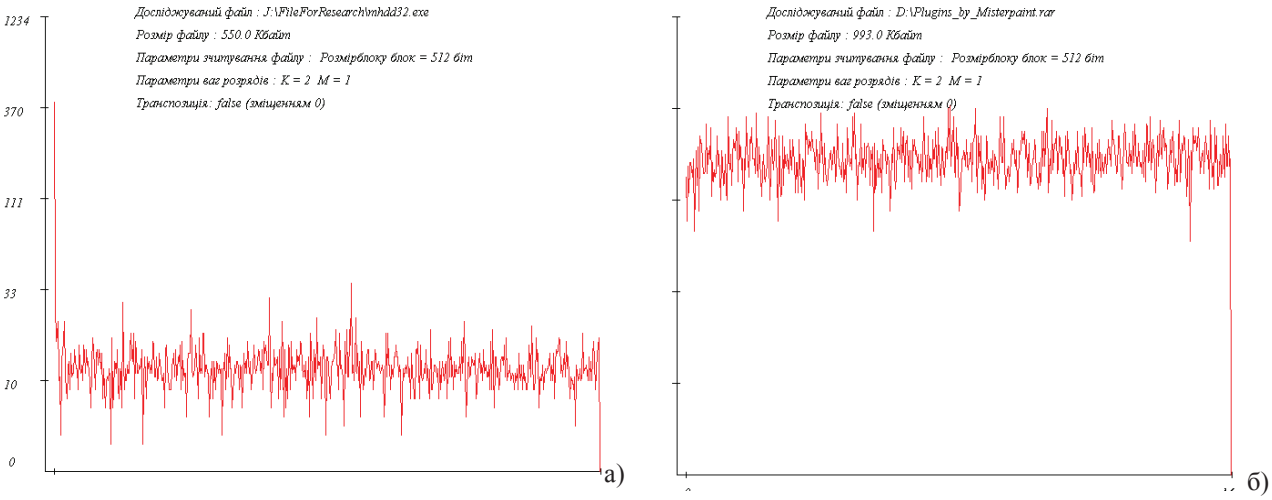
а) *.doc; б) *.pdf;

З отриманих результатів випливають такі особливості числових моделей для різних типів даних.

Для числової моделі файлу типу *.doc характерні:

- наявність екстремуму у першому піддіапізоні, що свідчить про те, що в числовій моделі є багато значень рівних або близьких нулю;
- наявність великої кількості чисел у середніх піддіапазонах;

- незначна кількість чисел у піддіапазонах, що близькі до максимального значення.



а) б)

Рисунок 9 – Результати досліджень моделей для різних типів даних:

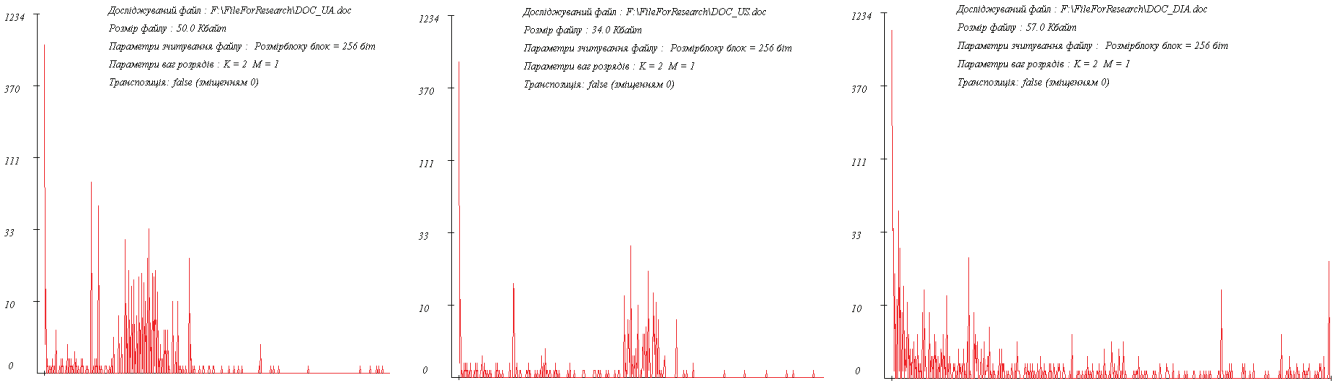
а) *.exe; б) *.zip

Числова модель файлу типу *.pdf характеризується рівномірним розподілом по всіх піддіапазонах.

Розподіл чисел у числовій моделі файлу типу *.exe має екстремум у першому піддіапізоні, що свідчить про те, що в числовій моделі є багато значень рівних або близьких нулю, а далі є рівномірним.

Для числової моделі файлу типу *.rar є характерним рівномірний розподіл по всіх піддіапазонах.

Також досліджувалось те, як на числову модель впливає вміст файлу. Дане дослідження важливе, тому що існує багато типів даних, які містять в собі інформацію різного типу. Наприклад, файли Microsoft office Document (DOC) можуть містити зображення, текст, формули, діаграми та інше. Залежність числової моделі від контенту файлу продемонстровано на рис. 10.



а) б) в)

Рисунок 10 – Результати досліджень моделей для файлів з різним вмістом:

а) *.doc з текстом на українській мові; б) *.doc з текстом на англійській мові; в) *.doc з діаграмами

Числова модель файлу типу Microsoft office Document (DOC) з текстом лише українською мовою має:

- екстремум у першому піддіапізоні, що свідчить про те, що в числовій моделі є багато значень рівних або близьких нулю;
- велику кількості чисел у піддіапазонах, що близькі до першого;
- незначна кількість чисел у піддіапазонах, що близькі до максимального значення.

Числова модель файлу типу Microsoft office Document (DOC) з текстом лише англійською мовою схожа до моделі файлу типу Microsoft office Document (DOC) з текстом лише українською мовою, але в ній велику кількість чисел у середніх піддіапазонах.

Висновки

Результати дослідження числових моделей файлів показали таке.

Розмір блоку, якому у відповідність ставиться число, значно вплива на статистичний розподіл. Із збільшенням розрядності він переходить від рівномірного до розподілу з характерними екстремумами.

Числова модель значно залежать від тип файлу. Модель файлу певного типу має свої особливості і здебільшого не схожа на моделі інших типів.

Значний вплив на числову модель файлу чинить його вміст. Файли одного і того ж типу, при різному вмісті, можуть мати різні числові моделі.

Урахування особливостей числових моделей забезпечить адаптацію алгоритму ущільнення для досягнення більшого степеня ущільнення.

Список використаної літератури

1. Балашов К.Ю. Сжатие информации: анализ методов и подходов. – Минск, 2000. – 42 с
2. Фомин А.А. "Основы сжатия информации" - Санкт-Петербург, 1998г.-289с.
3. Ватолін Д., Ракушняк А., Смірнов М., Юкин В. Методы сжатия данных. Устройство архиваторов, сжатие изображений и видео. – М.: ДИАЛОГ-ФИМИ, 2002. – 384с.
4. Bell T., Witten I, Cleary J. "Modeling for Text Compression", ACM Computing Surveys, Vol.21, No.4, pp.557-591, Dec. 1989.
5. Crochemore M., Lecroq Th. "Text data compression algorithms" Algorithms and Theory of Computation Handbook /Eds. M. J. Atallah, CRC Press Inc., Boca Raton, FL, 1998, pp.12.1-12.23.

Відомості про авторів

Лужецький Володимир Андрійович - д.т.н., професор, завідувач кафедри захисту інформації, Вінницький національний технічний університет, вул. Хмельницьке шосе 95, м.Вінниця.

Кульчицький Андрій Вікторович, магістрант, Вінницький національний технічний університет, вул. Хмельницьке шосе 95, м.Вінниця, тел.: (097)8844455, andriykulchitskiy@yandex.ru.

Алєксєєва Тетяна Михайлівна – студент, Вінницький національний технічний університет, вул. Хмельницьке шосе 95, м.Вінниця.