

TECHNICAL SCIENCES

ПОРІВНЯННЯ МЕТОДІВ ЕФЕКТИВНОГО КОДУВАННЯ ПОВІДОМЛЕНЬ

Розенвассер Д.М.

Одеська національна академія зв'язку ім. О.С. Попова

Педяш В.В.

Одеська національна академія зв'язку ім. О.С. Попова

COMPARISON OF MESSAGES EFFECTIVE CODING METHODS

Rozenvasser D.

O.S. Popov Odesa National Academy of Telecommunications

Pedyash V.

O.S. Popov Odesa National Academy of Telecommunications

АНОТАЦІЯ

Стаття присвячена вирішенню актуальної проблеми ефективного стиснення даних. Розглядаються існуючі методи стиснення даних, а також сучасні архіватори. На основі проведених досліджень та випробувань розроблено програмне забезпечення, яке надає можливість ефективного стиснення даних за допомогою використання найкращих можливостей з кожного розглянутого архіватора. Програма об'єднує сучасні архіватори та забезпечує ефективне використання ресурсів цих архіваторів для кожного типу файлів.

ABSTRACT

The article is devoted to solving the actual problem of effective data compression. Existing methods of data compression, as well as modern archivers are considered. Based on the research and testing carried out, software has been developed that makes it possible to effective data compression using the best capabilities for each archiver under consideration. The program combines modern archivers and ensures the efficient use of the resources of these archivers for each file type.

Ключові слова: кодування, повідомлення, стиснення даних, типи файлів, втрати, словник, архіватор.

Keywords: encoding, message, data compression, file types, loss, dictionary, archiver.

Завжди існувала і існує на сьогоднішній день проблема з розміром повідомлень при передаванні та зберіганні інформації. Розвиток комп'ютерного обладнання призвів до збільшення швидкості передачі та об'ємів зберігання даних. Може виникнути думка, що використання ефективних методів стиснення даних не потрібно на фоні цього технологічного розвитку. З іншого боку збільшується обсяг самих повідомлень. Збільшення обсягу повідомлень частіше відбувається швидше за збільшення швидкості передачі. Окрім цього люди повинні оплатити передачу інформації, навіть у фізичних носіїв є обмеження на кількість перезаписів. Саме ефективне стиснення може значно збільшити швидкість передачі інформації каналами зв'язку, збільшити ресурс твердих накопичувачів при відносно низьких затратах на придбання обладнання та програмного забезпечення. Також стиснення є також додатковою мірою захисту конфіденційної інформації під час її передачі. Так як зменшується кількість інформації, що передається, то зменшується вартість її передачі. Тому ефективне стиснення даних допомагає отримати перевагу у найрізноманітніших сферах, які пов'язані з даними.

Стиснення застосовується для зменшення кількості біт у порівнянні з джерелом, необхідних для зберігання або передачі заданої інформації. Це надає передачі інформації більшу швидкість та можливість її зберігати більш оперативно та економно. Оперативність означає, що операція отримання даної інформації з пристрою її зберігання проходить

швидше, а це можливо, якщо швидкість розпаковування даних вище за швидкість зчитування даних з носія інформації. Для стиснення інформації часто використовують програми-архіватори даних такі як ZIP, RAR, PA, 7ZIP та інші.

Для різних форматів даних використовують різні типи стиснення. Найчастіше для музики, відео та аудіо використовують стиснення з втратами, або незворотнього кодування, але для текстової інформації, фрагментів баз даних або файлів, які повинні виконуватися таке кодування призведе до втрати частини інформації, або навіть до неможливості коректної роботи цих файлів. Тому для цих видів даних використовують методи оборотнього стиснення, при яких гарантоване повне відновлення початкового повідомлення.

Мета даної роботи полягає у тому, щоб дослідити сучасні архіватори, створити алгоритм та розробити програмне забезпечення, яке дозволяє підвищити ефективність стиснення файлів шляхом обирання найбільш оптимального архіватора для конкретних видів інформації.

Стиснення та його види

Стиснення інформації - це проблема, яка має досить довгу історію, набагато довше, ніж історія розвитку комп'ютерних технологій, яка зазвичай супроводжувала історію розвитку проблеми кодування і шифрування інформації. Всі алгоритми стиснення працюють з вхідним потоком інформації, мінімальною одиницею якого є біт [1].

Стиснення даних засноване на усуненні надмірності, що міститься у вихідних даних:

- часові - в одновимірних даних, одновимірних сигналах, аудіо та інш.;
- просторові - кореляція між сусідніми пікселями або елементами даних;
- спектральні - кореляція між компонентами кольору або світла;
- психовізуальні - використовують перцептивні властивості зорової системи людини.

Найпростішим прикладом надмірності є повторення фрагментів в тексті. Така надмірність зазвичай усувається шляхом заміни послідовності, що повторюється посиланням на вже закодований фрагмент із зазначенням його довжини. Інший тип надмірності пов'язаний з тим, що деякі значення в стислих даних зустрічаються частіше, ніж інші. Скорочення обсягу даних досягається шляхом заміни даних, які часто зустрічаються короткими кодовими словами, а даних, які зустрічаються рідко – довгими (ентропійне кодування).

Стиснення без втрат даних, які не володіють властивістю надмірності (наприклад, випадковий сигнал або білий шум, зашифровані повідомлення), принципово неможливо.

Основними технічними характеристиками процесів стиснення і результатами їх роботи є:

- ступінь стиснення обсягів вхідного і вихідного потоків;
- ступінь стиснення - час, необхідний для стиснення певного обсягу інформації вхідного потоку, поки з нього не буде отриманий еквівалентний вихідний потік;
- якість стиснення - значення, що показує, наскільки упакований вихідний потік, шляхом повторного стиснення до нього з використанням того ж або іншого алгоритму.

Основною характеристикою алгоритму стиснення даних є коефіцієнт стиснення, тобто величина, яка визначає різницю між обсягом вихідних даних і стиснутих даних і може бути розрахована за наступною формулою:

$$k = \frac{S_0}{S_c} \quad (1)$$

де k – ступінь стиснення, S_0 – об'єм вихідних даних, S_c – об'єм даних, що стиснули.

Стиснення можливо поділити на дві категорії:

- 1) Стиснення з втратами;
- 2) Стиснення без втрат.

• Стиснення без втрат

Найпростіші алгоритми стиснення, також звані оптимальними алгоритмами кодування, є статистичними і засновані на розподілі ймовірностей елементів вхідного повідомлення (текст, зображення, файл). На практиці частота появи елементів у вхідному повідомленні використовується як наближення до ймовірності. Ймовірність – це абстрактна математична концепція, пов'язана з нескінченної експериментальної вибіркою даних, а частота виникнення – це значення, яке можна розрахувати для кінцевих наборів даних. При досить великій кількості елементів в наборі експериментальних даних можна сказати, що частота появи елемента близька (з деякою точністю) до його ймовірності.

Якщо згадані ймовірності різні, то можна зіставити коротші кодові слова для найбільш ймовірних елементів (загальні) і, навпаки, замінити елементи, які навряд чи зустрічаються, довгими кодовими словами.

Оптимальний алгоритм кодування робить це так, щоб середня довжина кодового слова була мінімальною, тобто, якщо довжина кодування коротше, воно стає незворотнім. Це робиться за допомогою префіксних алгоритмів Шеннона-Фано і Хаффмана.

Недолік обох методів полягає в тому, що вони не здатні кодувати повідомлення більш економічно, ніж один біт на елемент повідомлення (букву).

Арифметичне кодування є одним з алгоритмів ентропійного стиснення. На відміну від алгоритму Хаффмана, він не має жорсткого постійної відповідності між вхідними символами і групами бітів вихідного потоку. Це дає алгоритму велику гнучкість в поданні дрібних частот зустрічальності символів. Як правило, він перевершує алгоритм Хаффмана по ефективності стиснення і дозволяє стиснути дані з ентропією менше 1 біта на кодовий символ.

Існує інший фундаментальний підхід - словникове кодування, де на кожному кроці операції алгоритму стиснення поміщається наступний символ як є (зі спеціальним прапором без стиснення), або кодується слово з попереднього тексту, які відповідають наступним символам, які вказано у файлі. Розпакування стислих файлів здійснюється дуже швидко, тому ці алгоритми використовуються для створення програм, які саморозпаковуються. Словникові алгоритми менш математично обґрунтовані, але більш практичні.

Серед словників першим був алгоритм LZ77, розроблений ізраїльськими математиками Джейкобом Зивом і Абрахамом Лемпелем, опублікований в 1977 році. Багато програм стиснення використовують ту чи іншу модифікацію LZ77.

Стиснення з втратами

Алгоритми стиснення з втратами включають зменшення розміру файлу, зазвичай шляхом видалення дрібних деталей, які вимагають великого обсягу даних для зберігання з повною точністю. При стисненні з втратами неможливо відновити вихідний файл через вилучення важливих даних. Стиснення з втратами найчастіше використовується для зберігання зображень і звукової інформації.

Рекомендації CCITT G.711 пропонують два алгоритму для перетворення аудіосигналів: μ -закон і А-закон. Кодування μ -закону широко використовується в Сполучених Штатах і Японії, а кодування А-закону використовується в Європі. Обидва цих алгоритму перетворюють вихідні вибірки вихідної послідовності сигналів РСМ в байтові вибірки. Кожен зразок вихідної послідовності перетворюється в один байт.

Блочне кодування є одним з найпростіших методів кодування, які вимагають незначних обчислювальних витрат для стиснення зображень. Спочатку представлений Delp і Mitchell. Основна ідея BTC полягає в тому, щоб виконати квантування зі

збереженням моменту для блоків пікселів так, щоб якість зображення залишалася на задовільному рівні, і в той же час зменшувалася потреба в просторі для зберігання. Навіть якщо коефіцієнт стиснення алгоритму поступається стандартному алгоритму стиснення JPEG, BTC набув популярності завдяки своїй практичній корисності.

Дискретне косинусне перетворення (DCT) є найбільш популярним інструментом обробки сигналів для стиснення зображень і звуків, яке трапляється в таких стандартах, як JPEG.

Щоб виконати DCT-перетворення для зображення, спочатку потрібно отримати інформацію про фото зображення (значення пікселя у вигляді цілого числа, що має діапазон 0-255), яке ділимо на блоки матриці $N \times N$, а потім застосовуємо дискретне косинусне перетворення до цього блоку даних.

Дискретне перетворення має такі властивості:

- некорельовані коефіцієнти. Коефіцієнти не залежать одне від одного, тобто точність одного коефіцієнта не залежить від будь-якого іншого;
- енергія «ущільнення». Перетворення зберігає основну інформацію в невеликій кількості коефіцієнтів. Ця властивість найбільш помітна на фотореалістичних зображеннях.

Відео по суті являє собою тривимірний масив кольорових пікселів. Два виміру означають вертикальну та горизонтальну розширення кадру, а третій вимір - час. Кадр - це масив всіх пікселів, яких бачить камера в даний момент часу, або просто зображення.

Стиснення не було б можливим, якби кожен кадр був унікальним, а розташування пікселів було абсолютно випадковим, але це не так. Отже, ви можете стиснути, по-перше, саме зображення - наприклад, фотографія блакитного неба без сонця фактично зводиться до описом граничних точок і градієнта заповнення. По-друге, ви можете стиснути схожі суміжні кадри. В кінцевому рахунку, алгоритми стиснення зображень і відео схожі, якщо розглядати відео як тривимірне зображення з часом в якості третьої координати.

Однією з найбільш потужних технологій збільшення ступеня стиснення відеоданих є компенсація руху. У будь-якій сучасній системі стиснення відео наступні кадри в потоці використовують схожість областей в попередніх кадрах для збільшення ступеня стиснення.

Однак через рух будь-яких об'єктів в кадрі (або самої камери) використання подібності сусідніх кадрів виявилося неповним. Технологія компенсації руху дозволяє знаходити схожі області, навіть якщо вони зміщені щодо попереднього кадру.

Аналіз програм архівації даних

Використовуючи усі попередньо описані методи стиснення даних, а іноді й їх об'єднання, розроблені програмні засоби для ефективного стиснення різної інформації. У даній роботі було розглянуто 3 види архіваторів: 7Zip, WinRAR, PowerArchiver.

WinRAR - архіватор файлів для 32- і 64-розрядних операційних систем Windows (також існують або існували версії для Android, Linux, FreeBSD,

macOS, MS-DOS, Windows Mobile), що дозволяє створювати, змінювати і розпаковувати архіви RAR і ZIP, а також розпаковувати архіви безлічі інших форматів [2].

Особливості при роботі з архівами RAR:

- Підтримка файлів розміром до 16 ексабайт.
 - Розмір змінного словника від 1 МБ до 1 ГБ (в 32-розрядній версії для Windows - до 256 МБ). Розмір за замовчуванням - 32 МБ.
 - Замість застосовуваних за замовчуванням 32-розрядних контрольних сум CRC32 можна використовувати значно надійніше 256-розрядне хешування BLAKE2sp.
 - Можливість шифрування архівів з використанням алгоритму AES в режимі CBC з довжиною ключа 256 біт (у версії 4 - 128 біт).
 - Додавання в архіви додаткових, заснованих на кодах Ріда - Соломона, даних для відновлення архіву в разі його пошкодження, а також створення спеціальних томів для відновлення, що дозволяють відновити багатотомний архів при пошкодженні або навіть повній відсутності його окремих томів.
 - Додавання в архіви особливої додаткової інформації для прискорення їх відкриття.
 - Створення багатотомних (що складаються з декількох частин) архівів зазначеного або автоматично обирається розміру.
 - Створення безперервних (solid) архівів, що дозволяють досягати значно більш високого ступеня стиснення при упаковці декількох файлів, особливо однотипних.
 - Підтримка розширених можливостей NTFS, наприклад жорстких і символічних посилань.
- 7-Zip архіватор - вільний файловий архіватор з високим ступенем стиснення даних. Підтримує декілька алгоритмів стиснення і безліч форматів даних, включаючи власний формат 7z з високоефективним алгоритмом стиснення LZMA. Програма розробляється з 1999 року, вона безкоштовна і має відкритий вихідний код, велика частина якого вільно поширюється на умовах ліцензії GNU LGPL, за винятком коду Розпакувальник UnRAR, який має обмеження. Основна платформа - Windows (в тому числі Windows CE), де доступні дві версії програми: з графічним інтерфейсом і версія для командного рядка. Консольна версія була перенесена спільноту розробників для систем стандарту POSIX під загальною назвою p7zip.
- Основні характеристики 7-Zip:
- Дуже висока ступінь стиснення в форматі 7z завдяки використанню вдосконаленого алгоритму Лемпела-Зива.
 - Для форматів ZIP і GZIP ступінь стиснення на 2-10% вище, ніж у PKZIP і WinZip.
 - Можливість створення архівів для формату 7z.
 - Можливість створення багатотомних (порізаних на частини) архівів (за винятком саморозпаковуються) для формату 7z.

- Можливість шифрування алгоритмом AES-256 для форматів 7z і ZIP.
- Підтримує апаратне прискорення, доступне на процесорах, що реалізують набір інструкцій AES-NI.
- Інтеграція в оболонку Microsoft Windows і Windows NT.
- Модулі для програм FAR Manager, Total Commander і Unreal Commander.
- Вбудована утиліта для тестування продуктивності.
- Багатомовний графічний інтерфейс (тільки для Windows) з функціями двовіконну файлового менеджера.
- Потужна версія для командного рядка.

7-Zip використовує багатопоточність і дозволяє задіяти для стиснення, в залежності від алгоритму або формату, різну кількість потоків. При створенні архівів, в яких файли стискаються незалежно один від одного (наприклад ZIP), програма може використовувати до восьми потоків одночасно. Для алгоритму стиснення LZMA архіватор одночасно може використовувати до двох потоків. Неможливість використання більшої їх кількості пояснюється послідовним характером безперервного стиснення. Алгоритм стиснення LZMA2 не має цього недоліку [3].

При стисненні в форматі 7z також використовуються спеціальні фільтри-нормалізатори. Так, для більш оптимального стиснення 32-бітного x86-коду використовуються нормалізують конвертери BCJ і BCJ2. Крім того, програма має оптимізує дельта-конвертер для деяких типів мультимедійних даних, наприклад незжатих 24-бітних зображень. Зазвичай формат 7z дає на 4-25% краще стиснення, ніж формат ZIP.

PowerArchiver - універсальний архіватор з підтримкою всіх популярних форматів. Дозволяє робити з архівами найрізноманітніші дії: створення, перегляд, вилучення файлів; додавання файлів; створення архівів, які мають можливість саморозпаковки і багатотомних архівів; шифрування методами Blowfish 128-bit, DES 64-bit, Triple DES 128-bit і AES 128-bit; відновлення пошкоджених zip-архівів і багато іншого. Підтримує формати ZIP, RAR, 7-ZIP, CAB, LHA (LZH), TAR, TAR.GZ, TAR.BZ2, BH, ARJ, ARC, ACE, ZOO, GZ, BZIP2, XXE, MIME, UUE, XPI, EAR, WAR, REP, JAR, BK, QWK, Quake PK3, скіни WinAMP, архіви в форматах ZIP і CAB, CD-образи ISO, BIN, IMG і NRG. Можливе підключення антивірусної програми для перевірки архівних файлів [4].

Основні функції:

- Створення ZIP, CAB і PAE зашифрованих архівів за допомогою майстра.
- Підтримка необмеженого розміру ZIP архівів.
- Використання профілів стиснення для зберігання налаштувань компресії.
- Повна підтримка AES-шифрування з різним ступенем стійкості (128-bit / 192-bit / 256-bit).
- Повна підтримка AES-шифрування для 7-zip і багатотомних 7-Zip архівів.

– Завантаження архівів або резервних копій на FTP / SFTP (SSH) сайти з підтримкою профілів і функції тайм-ауту.

- Інтеграція з провідником Windows для швидкого створення та розпаковування архівів.
- Можливість редагування файлів в архіві, використовуючи одну операцію.
- Друк списку файлів в текстовий файл або файл формату HTML.
- Інструмент для відновлення пошкоджених архівів.
- Конвертація між архівами (наприклад, з ZIP в CAB або з LHA архіву в ZIP архів і т.д.).
- Одночасне витяг кількох архівів.
- Пакетний режим для створення декількох різних ZIP-архівів відразу.

Види файлів

У наш час широко використовуються файли самих різних типів. Серед них: doc, exe, bmp, pdf, wav, jpeg, mp3, xls, txt, html, tif. Деякі з них це «чисті» дані, тобто без використання методів стиснення до них, а деякі вже у стиснутому виді. Тому до останніх архівація діє набагато менше ніж на попередніх, тому що надлишковість у стиснутих файлах вже сильно зменшена.

ТХТ - текстові документи, збережені у форматі ТХТ можуть бути створені, відкриті і відредаговані з використанням різноманітних програм обробки редагування і слова тексту, розроблених для систем Linux, Microsoft Windows на комп'ютерах і Mac платформ. Зміст цих .txt файлів не відформатований ASCII текст, який може бути збережений як .txt документів в файли малого розміру [5].

JPEG - один з популярних растрових графічних форматів, який застосовують для зберігання фотозображень і подібних до них зображень. Файли даних JPEG, зазвичай мають розширення (суфікси) .jpg, .jif, .jpe або .jpeg. Однак з них .jpg є найпопулярнішим на всіх платформах. MIME-типом є image / jpeg [6].

Пакуються з використанням кодування серій і кодів Хаффмана. Стандарт JPEG допускає також використання значно більш ефективного арифметичного кодування, однак через патентні обмеження (патент на описаний в стандарті JPEG арифметичний QM-кодер належить IBM) на практиці воно використовується рідко.

EXE - розширення файлів, які виконуються, що застосовується в операційних системах DOS, Windows, Symbian OS, OS / 2 і в деяких інших, відповідне ряду форматів. Крім об'єктного коду може містити різні метадані (ресурси, цифровий підпис) [7].

Portable Document Format (PDF) - міжплатформений відкритий формат електронних документів, спочатку розроблений фірмою Adobe Systems з використанням низки можливостей мови PostScript. В першу чергу призначений для подання поліграфічної продукції в електронному вигляді. Для перегляду існує безліч програм, а також офіційна безкоштовна програма Adobe Reader [8,9].

MP3 - це розроблений командою MPEG формат файлу для зберігання аудіоінформації. MP3 є

одним з найбільш поширених і популярних форматів цифрового кодування звукової інформації. Він широко використовується в файлообмінних мережах для оціночного скачування музичних творів. У форматі MP3 використовується алгоритм стиснення з втратами, розроблений для істотного зменшення розміру даних, необхідних для відтворення запису і забезпечення якості відтворення звуку, що точно відповідає оригінальному (на думку більшості слухачів), але з відчутними втратами якості при прослуховуванні на якісній звуковій системі. Принцип стиснення полягає в зниженні точності деяких частин звукового потоку, що практично невиразно для слуху на найбільш поширеній апаратурі низької точності відтворення звуку (наприклад, домінуюча більшість портативних пристроїв, звукових карт, музичних центрів, автомагнітол та інших не спеціальної апаратури), а також для людей старшого віку, у зв'язку з природними віковими змінами слухового апарату, проте в більшості випадків чітко помітні на аудіотехніки високої точності відтворення звуку. Даний метод називають психовізуальним кодуванням. При створенні MP3 із середнім бітрейтом 128 кбіт/с в результаті виходить файл, розмір якого приблизно дорівнює 1/11 від оригінального файлу з CD-Audio (без стиснення бітрейт 1411,2 кбіт/с). MP3-файли можуть створюватися з високим або низьким бітрейтом, що впливає на якість файлу-результату [10].

BMP (англ. BitMap Picture) - апаратно-незалежне побітове зображення Windows - підтримується будь-якими Windows-сумісними програмами. Структура файлу BMP використовується Windows для зберігання растрових зображень. Наприклад, в цьому форматі зберігаються малюнки фону, піктограми та інші растрові зображення Windows. Формат зводить до мінімуму імовірність помилок або неправильної інтерпретації растрових даних. Основним недоліком даного формату є те, що тільки версії формату з 4- та 8-бітовим кольором піддаються стисненню. Отже, 24-бітові файли BMP будуть дуже великими. Крім того, застосування файлів BMP обмежене платформами Windows та OS/2. Все це робить неможливим застосування даного формату у мережі. У той же час цей формат використовується при створенні графічних файлів у графічних редакторах та при скануванні зображень. Кожен файл BMP містить заголовок файлу, заголовок зображення, растрові дані та карту кольорів (крім зображення 24-бітовим кольором). Заголовок файлу BMP містить інформацію про тип та розмір файлу, а також про розташування в ньому даних. Після заголовку файлу слідує структура, що задає інформацію про розмір, колір та стисненні зображення [11].

WAVE або WAV є короткою формою Wave Audio File Format (рідше іменованої як Аудіо для Windows). Цей формат є стандартом для зберігання аудіо потоку на ПК. Він є сферою докладання формату RIFF для зберігання аудіо в «ланцюжках», це дуже нагадує формати 8SVX та AIFF, використовувани комп'ютерами Amiga й Macintosh відповідно. Це також основний формат на системах Windows для зберігання звичайного нестиснутого звуку. Як правило, для цього застосовується кодування методом лінійної імпульсно-кодової модуляції [12].

Файл HTML відноситься до веб-сторінок, при створенні яких, використовувалася мова розмітки HTML. Зазвичай це не стиснений ASCII текст. Також може використовуватися UTF-8 [13].

Файл DOC (з англ. Microsoft Word Document) є документом, створеним в одному з найпопулярніших програмних комплексів для роботи з текстовими документами Microsoft Word. Варто відзначити, що будь-які версії пакетів Microsoft Office, включають в себе наявність програмного забезпечення Word. Формат DOC може зберігати в собі достатньо велику кількість різноманітної інформації, сюди входить різна інформація про форматування тексту, дані про вирівнювання, відступи, списки, абзаци і так далі. Подібні файли формату можуть містити не тільки текстовий документ з певною інформацією, але й всілякі зображення, таблиці, а також сценарії та діаграми [14].

Файл XLS - електронна таблиця, створена в Microsoft Excel - найпопулярнішому додатку для роботи з таблицями. Дані в документі зберігаються в осередках, кожна з яких має певну адресу [15].

Кожна клітинка може містити як фіксовані дані, так й формули, часто пов'язані з даними в інших осередках. Крім того, документ може містити зображення, а також діаграми, побудовані на основі даних у певних осередках.

TIFF (англ. Tagged Image File Format) - формат зберігання растрових графічних зображень. TIFF став популярним форматом для зберігання зображень з великою глибиною кольору. Він використовується при скануванні, надсилання факсимільних повідомлень, розпізнаванні тексту, у поліграфії, широко підтримується графічними додатками. TIFF був обраний в якості основного графічного формату операційної системи NeXTSTEP і з неї підтримка цього формату перейшла в Mac OS X. Формат був розроблений Aldus Corporation у співпраці з Microsoft для використання з PostScript [16]. Формат TIFF частіше інших використовується у поліграфії. Це найпопулярніший растровий формат для друку.

Випробування різних архіваторів

Таблиця 1

Випробування архіватора 7zip

Тип файлу	Розмір до стиснення, байт	Розмір після стиснення, байт	У скільки разів був стиснений файл відносно початкового	Коментарі
doc	2 450 944	198 127	~12.37	
exe	1 110 476	1 089 529	~1.01	немає стиснення
bmp	2 359 352	528 788	~4.46	
pdf	4 401 474	4 211 153	~1.04	немає стиснення
wav	609 964	480 438	~1.27	погане стиснення
jpeg	775 702	772 879	~1.00	немає стиснення
mp3	8 414 449	8 259 672	~1.01	немає стиснення
xls	2 054 144	346 622	~5.9	
txt	848 410	294 929	~2.87	
html	400 444	71 708	~5.58	
tif	101 098	68 829	~1.47	

Таблиця 2

Випробування архіватора Zip

Тип файлу	Розмір до стиснення, байт	Розмір після стиснення, байт	У скільки разів був стиснений файл відносно початкового	Коментарі
doc	2 450 944	312 192	~7.8	
exe	1 110 476	1 092 954	~1.01	немає стиснення
bmp	2 359 352	675 844	~3.49	
pdf	4 401 474	4 285 599	~1.02	немає стиснення
wav	609 964	571 114	~1.06	немає стиснення
jpeg	775 702	772 482	~1.00	немає стиснення
mp3	8 414 449	8 294 635	~1.01	немає стиснення
xls	2 054 144	533 632	~3.85	
txt	848 410	340 239	~2.50	
html	400 444	90 172	~4.44	
tif	101 098	70 130	~1.44	

Таблиця 3

Випробування архіватора WinRAR

Тип файлу	Розмір до стиснення, байт	Розмір після стиснення, байт	У скільки разів був стиснений файл відносно початкового	Коментарі
doc	2 450 944	211 573	~11.58	
exe	1 110 476	1 093 377	~1.01	немає стиснення
bmp	2 359 352	561 769	~4.20	
pdf	4 401 474	4 241 040	~1.03	немає стиснення
wav	609 964	489 983	~1.24	погане стиснення
jpeg	775 702	773 637	~1.00	немає стиснення
mp3	8 414 449	8 260 582	~1.01	немає стиснення
xls	2 054 144	442 205	~4.64	
txt	848 410	310 000	~2.73	
html	400 444	78 403	~5.10	
tif	101 098	68 912	~1.46	

Таблиця 4

Випробування архіватора Power archiver (Extreme parameters)

Тип файлу	Розмір до стиснення, байт	Розмір після стиснення, байт	У скільки разів був стиснений файл відносно початкового	Коментарі
doc	2 450 944	182 804	~13.40	
exe	1 110 476	1 093 715	~1.01	немає стиснення
bmp	2 359 352	486 335	~4.85	
pdf	4 401 474	3 306 264	~1.33	погане стиснення
wav	609 964	479 009	~1.27	погане стиснення
jpeg	775 702	630 344	~1.23	погане стиснення
mp3	8 414 449	-----	-----	помилка
xls	2 054 144	324 898	~6.32	
txt	848 410	245 502	~3.45	
html	400 444	51 441	~7.78	
tif	101 098	68 158	~1.48	

Таблиця 5

Випробування архіватора Power archiver (Normal parameters)

Тип файлу	Розмір до стиснення, байт	Розмір після стиснення, байт	У скільки разів був стиснений файл відносно початкового	Коментарі
doc	2 450 944	199 829	~12.2	
exe	1 110 476	1 090 685	~1.02	немає стиснення
bmp	2 359 352	528 044	~4.47	
pdf	4 401 474	3 317 254	~1.33	погане стиснення
wav	609 964	481 422	~1.27	погане стиснення
jpeg	775 702	637 750	~1.22	погане стиснення
mp3	8 414 449	-----	-----	помилка
xls	2 054 144	351 038	~5.85	
txt	848 410	245 502	~3.45	
html	400 444	51 441	~7.78	
tif	101 098	68 566	~1.47	

Порівнюємо усі раніше отримані результати, та оберемо найкращі коефіцієнти стиснення.

Таблиця 6

Порівняння архіваторів 7zip, WinRar та PA

Тип файлу	Рекомендований архіватор	Розмір до стиснення, байт	Розмір після стиснення, байт	У скільки разів був стиснений файл відносно початкового
doc	7zip	2 450 944	198 127	~12.37
exe	PA	1 110 476	1 090 685	~1.02
bmp	PA	2 359 352	528 044	~4.47
pdf	PA	4 401 474	3 317 254	~1.33
wav	7zip	609 964	480 438	~1.27
jpeg	PA	775 702	637 750	~1.22
mp3	7zip	8 414 449	8 259 672	~1.01
xls	7zip	2 054 144	346 622	~5.9
txt	PA	848 410	245 502	~3.45
html	PA	400 444	51 441	~7.78
tif	PA	101 098	68 566	~1.47

Розробка програми

На основі отриманих даних була розроблена програма для оптимальної архівації. Програма була розроблена на мові C#. Для кожного файлу який користувач переміщує у неї, вона обирає оптимальний архіватор та відкриває його для користувача. Якщо програмі не було визначено тип файлу, який користувач бажає заархівувати, то вона за замовчуванням пропонує використовувати WinRAR так як він підтримує саму більшу різноманітність файлів, які можуть за допомогою нього бути заархівовані.

Також для файлів типу txt програма пропонує на вибір 2 типи можливих архіваторів: арифметичне кодування (найбільш оптимальне рішення) та Power arch (як більш зручний у використанні). Для типів файлів exe, bmp, pdf, jpeg, html, tif програмою рекомендований архіватор Power arch, для doc, wav, mp3, xls рекомендований архіватор 7zip. При надходженні невідомого типу файлу було вирішено використовувати WinRAR так як він має саму велику кількість файлів, що може підтримувати.

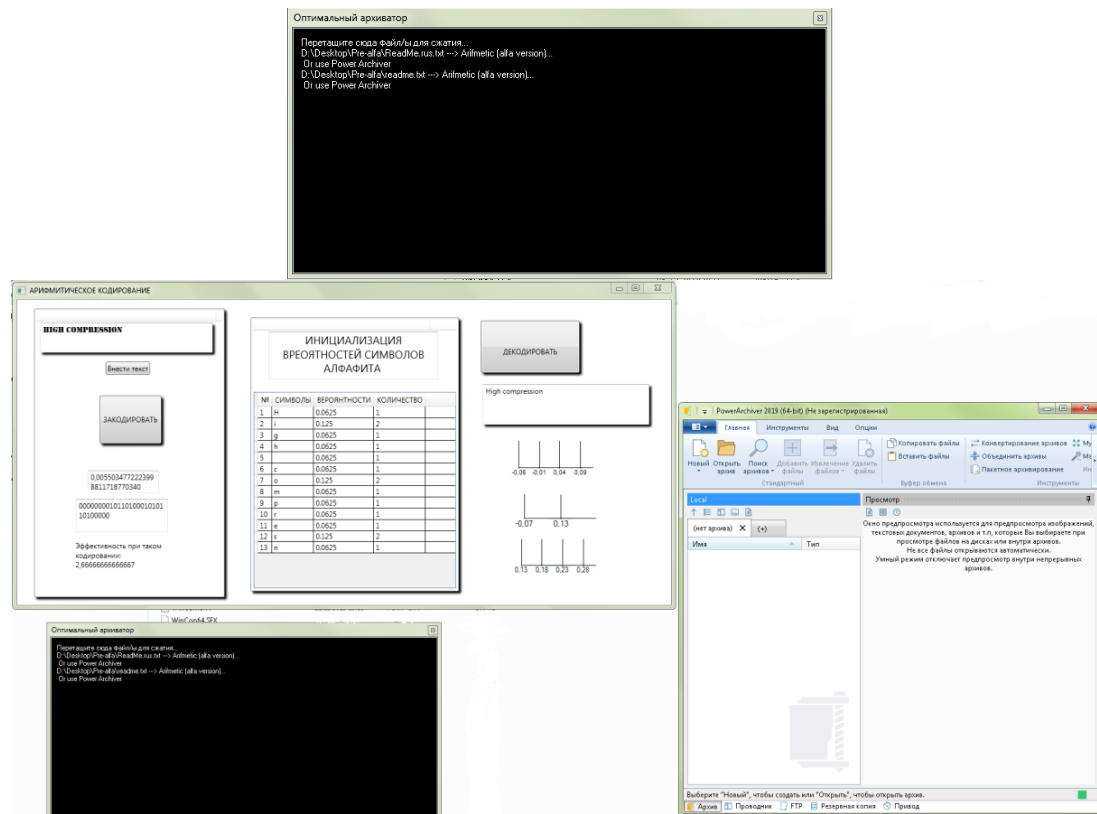


Рис. 1. Демонстрація роботи програми оптимальної архівації

Висновки. У сучасний час файли досягли дуже великих обсягів, розмір файлів зростає швидше ніж можливості носіїв. Тому були розроблені програми архіватори, які могли без втрати даних зменшити обсяг необхідний для їх збереження.

На даний час пропонують використовувати різні архіватори. Кожен архіватор має деякі переваги над іншими, але не існує архіватора, який був би у всьому набагато кращим за своїх конкурентів.

Тому у даній роботі було проведено дослідження того, який коефіцієнт стиснення найкращий у кожного архіватора до певних типів файлів. З кожного архіватора було обране найкраще й об'єднано до програми, яка допомагає обирати найоптимальніший архіватор до конкретних типів файлів щоб досягти найкращого результату по місцю, який займає файл після обробки.

Література

1. Data compression <https://searchstorage.techtarget.com/definition/compression> (дата звернення 11.11.2020).
2. WinRAR: <https://www.winrar.com/education.html> (дата звернення 11.11.2020).
3. 7-Zip: <https://www.7-zip.org/> (дата звернення 11.11.2020).
4. PowerArchiver: <https://www.powerarchiver.com/> (дата звернення 11.11.2020).
5. .TXT Расширение файла: <https://www.reviversoft.com/ru/file-extensions/txt> (дата звернення 11.11.2020).

6. JPEG: <https://uk.wikipedia.org/wiki/JPEG> (дата звернення 11.11.2020).

7. .EXE: <https://www.online-convert.com/file-format/exe> (дата звернення 11.11.2020).

8. Параметры Adobe PDF: <https://helpx.adobe.com/ru/indesign/using/pdf-options.html> (дата звернення 11.11.2020).

9. Portable Document Format: https://ru.wikipedia.org/wiki/Portable_Document_Format (дата звернення 11.11.2020).

10. MP3: <https://uk.wikipedia.org/wiki/MP3> (дата звернення 11.11.2020).

11. BMP (Bitmap Picture): [https://ru.bmstu.wiki/BMP_\(Bitmap_Picture\)](https://ru.bmstu.wiki/BMP_(Bitmap_Picture)) (дата звернення 11.11.2020).

12. Формат WAV: <https://audiocoding.ru/formats/wav/> (дата звернення 11.11.2020).

13. Файл формата HTML — что это?: <https://filesreview.com/ru/info/html> (дата звернення 11.11.2020).

14. Файл формата DOC — что это?: <https://filesreview.com/ru/info/doc> (дата звернення 11.11.2020).

15. Файл с расширением .xls: <https://open-file.ru/types/xls> (дата звернення 11.11.2020).

16. Формат файлов TIFF в деталях: http://docscan.ru/allabout/tiff_file_format.html (дата звернення 11.11.2020).