УДК 004.63

ПОРІВНЯННЯ ЕФЕКТИВНОСТІ СТИСНЕННЯ ЕЛЕКТРОННИХ КНИГ У РІЗНИХ ФОРМАТАХ

Захаров В.В.
email: volodymyr.zakharov@nure.ua
Науковий керівник – д.т.н., проф. Машталір С.В.
Харківський національний університет радіоелектроніки, каф. ІНФ
м. Харків, Україна

This work is devoted to assessing the efficiency of e-book compression in various formats. A comparative analysis of EPUB, MOBI, and PDF formats was conducted based on file size reduction for different literary works. The study demonstrated that EPUB provides the most efficient compression, consistently producing the smallest file sizes. MOBI exhibits moderate compression efficiency, while PDF results in significantly larger files due to its document-oriented structure. The findings highlight EPUB as the optimal format for minimizing storage requirements.

У сучасному цифровому середовищі електронні книги стали невід'ємною частиною повсякденного життя. Однак розмір файлів електронних книг може значно відрізнятися залежно від використовуваного формату. Дослідження ефективності стиснення електронних книг у різних форматах є актуальним, оскільки дозволяє визначити оптимальні рішення для зменшення обсягу файлів без суттєвих втрат якості тексту та графічного контенту.

У цій роботі проведено порівняльний аналіз розміру файлів у форматах EPUB, MOBI та PDF для різних літературних творів. Отримані результати допоможуть визначити найбільш ефективний формат з точки зору компресії та збереження читабельності документа.

Файл PDF – це індексована колекція об'єктів. Об'єкт PDF – це частина структурованих даних. PDF-файл складається із заголовка, купи визначень об'єктів, таблиці перехресних посилань і трейлера. Таблиця перехресних посилань – це таблиця пошуку, яка надає розташування кожного пронумерованого об'єкта як зміщення байтів у файлі. Трейлер містить інформацію про кореневий об'єкт або каталог документів, який є відправною точкою для інтерпретації файлу PDF. У кінці файлу є зміщення байтів до таблиці перехресних посилань.

Для того, щоб ці зсуви байтів були корисними, PDF-файл повинен бути доступним для пошуку, тобто читач повинен мати можливість прочитати частину файлу, починаючи з заданого зміщення байтів у файлі. Формат файлу PDF створено таким чином, що ви можете одразу відобразити ту частину файлу, яка вас цікавить, без необхідності читати частини, які вас не цікавлять. Це одна з найважливіших цілей формату PDF.

EPUB (electronic publication) – це відкритий стандарт Міжнародного форуму цифрових видавців (IDPF) для створення та розповсюдження цифрових публікацій, наприклад електронних книг. Вміст EPUB є «перекомпонованим», що означає, що до нього можна отримати доступ на будь-якому з численних пристроїв для читання електронних книг, які підтримують стандарт (Kindle, Sony Reader, Nook, Kobo тощо), а також на більшості смартфонів і планшетів.

Документ EPUB складається з OPF, XML, XHTML, HTML, CSS, NCX та файлів зображення у єдиному сумісному форматі файлів, скомпресованих за допомогою ZIP, для легкого розповсюдження та публікації. Внутрішню структури EPUB можна легко перевірити, відкривши файл .epub за допомогою файлового архіватору, наприклад WinRAR (рис. 1)

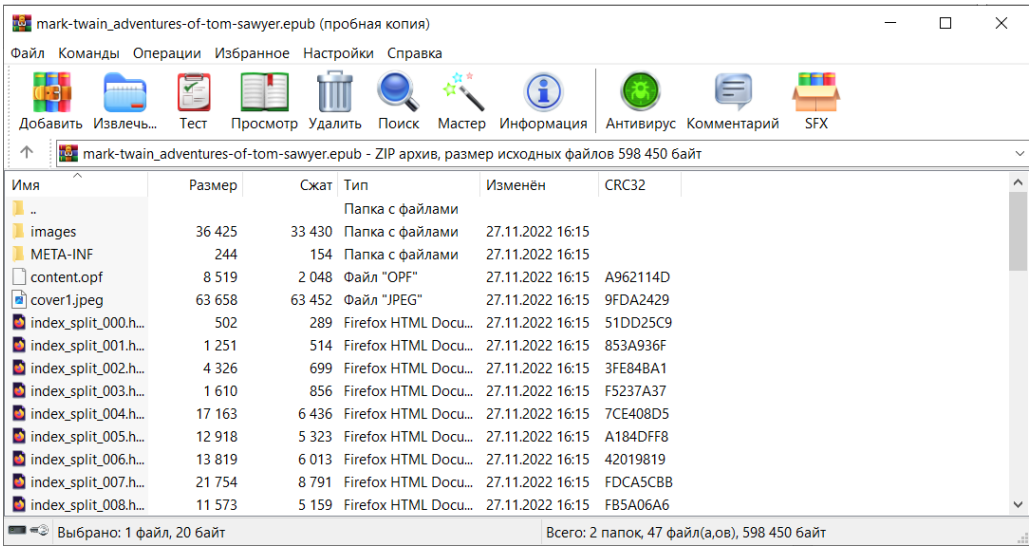

Рисунок 1 – Відкриття mark-twain_adventures-of-tom-sawyer.epub у WinRar

Формат MOBI спочатку був розширенням формату PalmDOC шляхом додавання певних HTML-тегів до даних (див. EBook HTML). Багато документів у форматі MOBI все ще використовують цю форму. Однак існує також версія цього формату файлу з високим рівнем стиснення, яка стискає дані більшою мірою запатентованим способом. Режим вищого стиснення використовує схему кодування Хаффмана, яку називають Huff/cdic алгоритм.

PalmDOC використовує методи стиснення LZ77, реалізацію для PalmDOC можна знайти на Github. Файли DOC можуть містити лише стиснутий текст. Формат не допускає жодного форматування тексту. Завдяки цьому файли залишаються невеликими відповідно до філософії Palm.

Алгоритми LZ77 досягають стиснення шляхом заміни частин даних посиланнями на відповідні дані, які вже пройшли через кодер і декодер. Збіг

кодується парою чисел, яка називається парою довжина-відстань, що еквівалентно вислову «кожен із наступних символів довжини дорівнює символу, що знаходиться на точній відстані від нього в нестисненому потоці».

У форматі PalmDoc пара довжина-відстань завжди кодується двобайтовою послідовністю. З 16 біт, які складають ці два байти, 11 біт йдуть на кодування відстані, 3 йдуть на кодування довжини, а решта два використовуються, щоб переконатися, що декодер може визначити перший байт як початок такої двобайтової послідовності. Дані PalmDOC завжди поділяються на блоки по 4096 байт (розмір без стиснення), і блоки обробляються незалежно; під час стиснення або розпакування блоку інформація з попередніх чи наступних блоків не потрібна.

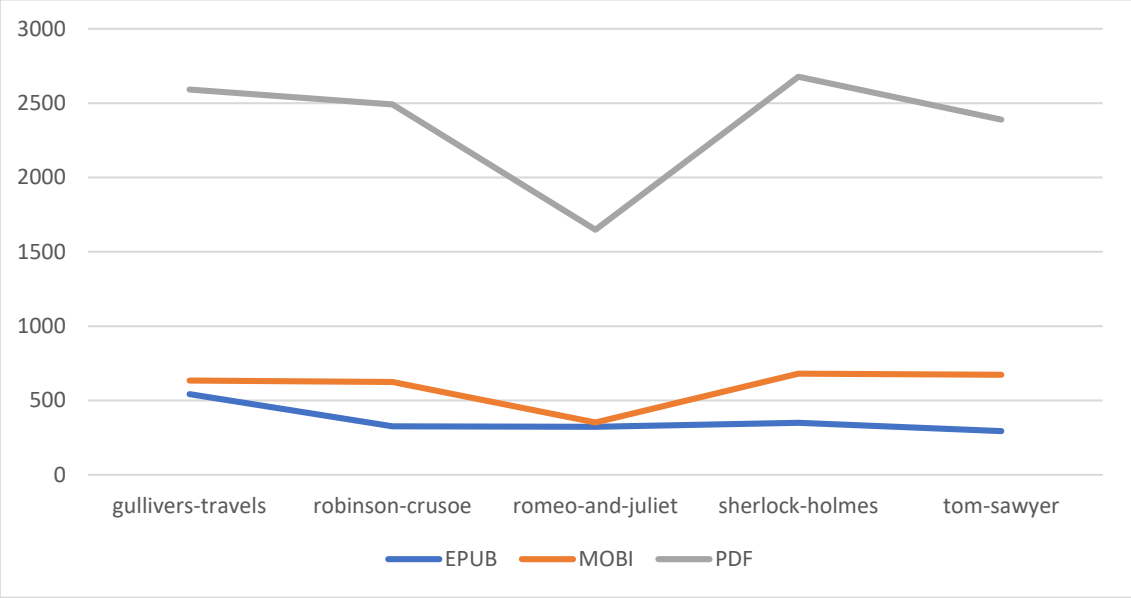

Рисунок 2 – Порівняння ваги книг у різних форматах

В сумі: як можна побачити на рис. 2, формат EPUB стискає файли найкраще, за що і був обраний стандартом IDPF, стиснення MOBI майже дорівнює EPUB, але PDF стискає електронні книги набагато гірше, бо PDF є документ-орієнтованим форматом.

Список використаних джерел:

1. https://www.globalgreyebooks.com/adventures-of-sherlock-holmes-ebook.html
2. https://www.globalgreyebooks.com/robinson-crusoe-ebook.html
3. https://www.globalgreyebooks.com/romeo-and-juliet-ebook.html
4. https://www.globalgreyebooks.com/adventures-of-tom-sawyer-ebook.html
5. https://www.globalgreyebooks.com/gullivers-travels-ebook.html