УДК 004.627

В. О. Сірак[1]
А. В. Снігур[1]

ПОРІВНЯЛЬНИЙ АНАЛІЗ АЛГОРИТМІВ СТИСКАННЯ ДАНИХ НА ПРИКЛАДІ МЕТОДІВ ХАФМЕНА ТА LZW

[1]Вінницький національний технічний університет

Анотація

Розглянуто принципи роботи алгоритмів стискаючого кодування даних без втрат Хафмена та LZW. На основі наукових публікацій порівняно їхні особливості, переваги й обмеження під час оброблення текстових, графічних та аудіоданих. Визначено критерії оцінювання ефективності стискання та обґрунтовано залежність результатів від характеристик вхідної інформації.

Ключові слова: стискання даних, кодування без втрат, алгоритм Хафмена, LZW, коефіцієнт стискання.

Abstract

The paper examines the operating principles of Huffman and LZW lossless data compression algorithms. Their features, advantages, and limitations for processing text, image, and audio data are compared based on scientific publications. The main compression efficiency criteria are identified, and the dependence of results on input data characteristics is discussed.

Keywords: data compression, lossless coding, Huffman algorithm, LZW, compression ratio.

Вступ

Збільшення обсягів цифрової інформації зумовлює потребу в ефективних методах її зберігання та передавання. Одним зі способів зменшення обсягу даних є стискання без втрат, яке забезпечує повне відновлення початкової інформації після декодування.

Серед таких методів важливе місце посідають алгоритм Хафмена, що використовує коди різної довжини залежно від частоти появи символів, та LZW (Lempel–Ziv–Welch), який ґрунтується на словниковому кодуванні повторюваних послідовностей [1].

Метою роботи є порівняння алгоритмів Хафмена та LZW за принципами роботи, особливостями застосування й ефективністю стискання різних типів даних.

Результати дослідження

Алгоритм Хафмена належить до методів статистичного кодування. На основі частот появи символів будується бінарне дерево, за допомогою якого формуються кодові слова: часті символи зазвичай отримують коротші коди, а рідкісні - довші. Префіксна властивість кодів забезпечує однозначне декодування. Водночас для відновлення даних необхідно зберігати або передавати інформацію про відповідні коди.

Алгоритм LZW використовує словник послідовностей символів, який формується під час кодування. Знайдені послідовності замінюються відповідними кодами, а нові комбінації додаються до словника. Такий підхід дає змогу ефективно використовувати повторювані фрагменти даних без попереднього обчислення частот усіх символів.

Для порівняння алгоритмів доцільно застосовувати коефіцієнт стискання:

$$K = \frac{V_0}{V_c},$$

де (V_0) - початковий обсяг даних, а (V_c) - обсяг після стискання. Що більше значення (K), то сильніше стиснено дані. Під час оцінювання необхідно враховувати службову інформацію, потрібну для декодування, а також час виконання обох операцій.

У дослідженні М. Н. Фаузана, М. Аліфа та К. Пріанто безпосередньо порівняно алгоритми Хафмена та LZW під час стискання текстових файлів. Автори оцінювали можливість зменшення обсягу файлів, що робить цю роботу релевантною для аналізу ефективності двох методів на текстових даних [1].

Д. Клен, Й. Лерга та І. Петрієвчанін розглянули застосування LZW, Хафмена й арифметичного кодування для текстових файлів і зображень. У роботі використовувалися показники ступеня

стискання та часу виконання. Це підтверджує важливість оцінювання алгоритмів не за одним параметром, а за сукупністю характеристик і з урахуванням типу даних [2].

Інший аспект розглянуто в дослідженні С. Шукли та співавторів, присвяченому стисканню аудіосигналів. Автори порівнювали кілька методів стискання без втрат, зокрема Хафмена та LZW. Ця робота розширює порівняння за межі текстових файлів і демонструє доцільність перевірки алгоритмів на даних із різними властивостями [3].

Кількісні результати попередніх досліджень підтверджують залежність ефективності стискання від типу даних. У дослідженні М. Н. Фаузана, М. Аліфа та К. Пріанто для алгоритму LZW наведено середню економію місця 63,85 % для TXT-файлів і 77,56 % для CSV-файлів [1]. У роботі Д. Клена, Й. Лерги та І. Петрієвчанін наведено медіанні показники стискання для текстових даних і зображень: близько 30 % і 70 % для LZW та 40 % і 55 % для алгоритму Хафмена відповідно [2]. Оскільки в цих дослідженнях застосовувалися різні показники оцінювання, числові результати не слід безпосередньо зіставляти між собою. Вони свідчать про залежність ефективності стискання від характеристик вхідних даних.

Узагальнену характеристику методів наведено в таблиці 1.

Таблиця 1 - Порівняння алгоритмів Хафмена та LZW

Критерій	Хафмена	LZW
Принцип роботи	Кодування з урахуванням частоти символів	Словникове кодування послідовностей
Основна перевага	Короткі коди для частих символів	Повторне використання кодів відомих послідовностей
Особливість декодування	Потрібна таблиця кодів або дерево	Словник відновлюється під час декодування
Чинник ефективності	Розподіл частот символів	Повторюваність послідовностей
Критерії оцінювання	Обсяг результату та час виконання	Обсяг результату та час виконання

Висновки

Порівняльний аналіз показав, що алгоритми Хафмена та LZW використовують різні принципи стискання, а їхня ефективність залежить від структури вхідних даних. Розглянуті наукові дослідження підтверджують, що жоден із цих методів не є універсально найефективнішим для всіх типів інформації. Тому під час вибору алгоритму доцільно враховувати коефіцієнт стискання та час виконання. Отримані результати можуть бути використані для обґрунтованого вибору методу стискання залежно від типу даних.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Fauzan M. N., Alif M., Prianto C. Comparison of Huffman Algorithm and Lempel Ziv Welch Algorithm in Text File Compression. *IT Journal Research and Development*. 2022. Vol. 7, No. 2. P. 184–197. DOI: https://doi.org/10.25299/itjrd.2023.10437.
2. Klen D., Lerga J., Petrijevčanin I. Txt and Tif File Compression Using LZW, Huffman, and Arithmetic Coding. *Pomorski zbornik*. 2024. Special edition 5. P. 213–228. URL: https://hrcak.srce.hr/317229.
3. Shukla S., Gupta R., Rajput D. S., Goswami Y., та ін. A Comparative Analysis of Lossless Compression Algorithms on Uniformly Quantized Audio Signals. *International Journal of Image, Graphics and Signal Processing*. 2022. Vol. 14, No. 6. P. 59–69. DOI: https://doi.org/10.5815/ijigsp.2022.06.05.

Науковий керівник: ***Снігур Анатолій Васильович*** – кандидат технічних наук, доцент кафедри обчислювальної техніки, Вінницький національний технічний університет, Вінниця, e-mail: snigur.anatoliy@vntu.edu.ua

Сірак Владислав Олегович – студент групи 1КІ-25мс2, факультет інформаційних технологій та комп'ютерної інженерії, Вінницький національний технічний університет, Вінниця, e-mail: vladsirak9@gmail.com

Supervisor: ***Snigur Anatolii Vasylovych*** – Candidate of Technical Sciences, Associate Professor of the Computer Engineering Department, Vinnytsia National Technical University, Vinnytsia, e-mail: snigur.anatoliy@vntu.edu.ua

Sirak Vladislav Olehovych – student of group 1KI-25ms2, Faculty of Information Technologies and Computer Engineering, Vinnytsia National Technical University, Vinnytsia, e-mail: vladsirak9@gmail.com