УДК 004.21

ДОСЛІДЖЕННЯ АЛГОРИТМІВ ОПТИМАЛЬНОГО КОДУВАННЯ І СТИСНЕННЯ ДАНИХ

Мичка С. О.
Науковий керівник – к.т.н., доцент Голян Н. В.
Харківський національний університет радіоелектроніки, каф. ПІ,
м. Харків, Україна
e-mail: sviatoslav.mychka@nure.ua

File compression algorithms have long been integral to information theory, serving diverse purposes in data storage, transmission, and resource optimization. With the proliferation of the Internet, their significance has only heightened across various domains. This paper describes the research on compression algorithms, which employ increasingly sophisticated techniques to enhance compression efficiency across diverse data types. Specifically, this research aims to analyze and compare the behavior of prominent compression algorithms, including entropy encodings, dictionary methods, and context modelling, with respect to their effectiveness in compressing data types.

У сучасному світі кількість нових даних і, відповідно, затрати на їх зберігання й транспортування таких об’ємів даних теж зростають. Актуальність розробки та поліпшення алгоритмів стиснення даних полягає, насамперед, у тому, що сильніше стиснення даних із можливістю точно (за використання алгоритмів стиснення без втрат) відновити початкову інформацію дозволяє зменшити витрати на їх зберігання, транспортування й обробку. Алгоритми стиснення без втрат можуть бути ентропійними, словниковими, або поєднувати обидва підходи, а також використовувати інші допоміжні методи, такі як перетворення вихідного повідомлення або моделювання контексту. Тому деякі алгоритми можуть давати більші коефіцієнти стиснення з одними даними, і менші – з іншими, що потенційно може призвести до зайвих витрат ресурсів на зберігання або передавання даних.

Для вирішення цієї проблеми було поставлено задачу дослідження методів оптимального кодування і стиснення даних і порівняння характеристик їхньої роботи, знаходження алгоритмів стиснення, що працюють краще на певних видах даних. Потрібно дослідити алгоритми, їхні поєднання, підходи та методи стиснення без втрат даних різних видів, визначити критерії порівняння тестових даних та алгоритмів, що досліджуються, і використати їх для порівняння ефективності застосування методів стиснення до певних видів файлів. Використовуючи отримані дані, необхідно розробити додаток, що підбиратиме необхідний алгоритм стиснення для даних, наданих користувачем, а також провести ряд експериментів для оцінки ефективності роботи розробленого додатку.

Було проведено аналіз та обрано для дослідження алгоритми словникового стиснення даних LZ77, LZW і RLE із використанням перетворення Барроуза-Віллера, ентропійного стиснення за допомогою коду Хафмана, поєднання ентропійного та словникового методів у вигляді алгоритму Deflate, а також алгоритми, що застосовують моделювання контексту: PPM і Brotli. Ці алгоритми в наш час отримують розвиток і популярність в першу чергу через необхідність передавати більші об'єми даних мережею Інтернет і зберігати їх у хмарних сховищах.

Було розроблено план проведення експериментів, згідно з яким:

– визначено та поділено на категорії вхідні дані, що використовуватимуться в якості тестових, а саме: за середньою ентропією повідомлення, за наближеністю до тексту натуральною мовою, за повторюваністю символів;

– виконано стиснення тестових даних за допомогою алгоритмів, використовуючи бібліотеки або власні реалізації;

– порівняно результати, використовуючи такі метрики: коефіцієнт стиснення, час кодування, час декодування, кількість використаної оперативної пам'яті;

– розроблене програмне забезпечення для надання рекомендацій щодо використання алгоритмів стиснення для різних даних, використовуючи дані, отримані під час стиснення тестових даних;

– порівняно рекомендації щодо тестових даних із фактичними результатами їх стиснення.

Експерименти планується проводити з використанням локального комп'ютера з 4-ядерним процесором Intel Core i7-7700HQ і 16 Гб оперативної пам'яті (DDR4). Очікується, що метрики, отримані під час проведення експериментів дозволять краще зрозуміти принципи роботи алгоритмів із реальними даними. Вони можуть бути корисними для подальших досліджень в області стиснення даних, а також для розробки й покращення практичних рішень, що вже існують. Розроблене програмне забезпечення не лише може бути використане як практична реалізація для демонстрації результатів досліджень, а й потенційно має застосування як API для визначення оптимального алгоритму стиснення даних для їхнього подальшого зберігання або передачі мережею Інтернет.

Список використаних джерел:

1. David J. C. MacKay. (2005). Information Theory, Inference, and Learning Algorithms. Cambridge University Press.

2. Sharonova N., Kyrychenko I., Gruzdo I., Tereshchenko G. (2022). Generalized Semantic Analysis Algorithm of Natural Language Texts for Various Functional Style Types. CEUR Workshop Proceedings, 3171, 16–26. https://ceur-ws.org/Vol-3171/paper4.pdf.