УДК 681.32

Д. О. Пущал

ПАРАЛЕЛЬНА ОБРОБКА АЛГОРИТМІВ УЩІЛЬНЕННЯ ТЕКСТОВИХ ТА ЧИСЛОВИХ ДАНИХ

Вінницький національний технічний університет

Анотація *Розглянуто алгоритми ущільнення даних. Виділено статистичний та адаптивний алгоритм Хаффмана. Проаналізовано доцільність використання алгоритмів для досягнення задачі по паралельному ущільненні текстової інформації.*

Ключові слова: ущільнення, паралельна обробка, метод Хаффмана.

Abstract *Data compression algorithms are considered. The statistical and adaptive Huffman algorithm is distinguished. The expediency of using algorithms to achieve the problem of parallel compression of text information is analyzed.*

Keywords: compaction, parallel processing, Huffman's algorythm.

Для ефективного кодування повідомлень використовують теорему Шеннона, яку формують так:

- повідомлення джерела з ентропією H(z) завжди можна закодувати послідовностями символів з об'ємом алфавіту m так, що середнє число символів на знак повідомлення $l_{ср}$ буде як завгодно близьким до величини H(z)/log(m), де m=2, $l_{ср} \geq H(z)$. але не менше за неї. [1]

Кожен символ вихідної комбінації містить максимум інформації, яку записано двійковим кодом.

Постала проблема ефективного використання пам'яті, що зумовило необхідність створення спеціальних алгоритмів та засобів роботи для ущільнення інформації які реалізовані в програмах-архіваторах.

Два основних напрями ущільнення – lossy і lossless. Перша група включає в себе методи кодування з втратами, друга – методи ущільнення без втрат, яка є більш універсальними.

Ущільнення – спосіб ефективно використовувати пам'ять зменшуючи розмір вихідних файлів за допомогою спеціально створених для цієї задачі алгоритмів.

Сучасні машини кожного дня оперують великими об'ємами даних, і для заощадження місця на фізичних носіях, на яких зберігаються результуючі дані, використовують методи ущільнення. З урахуванням того, що сучасні ЕОМ «озброєні» процесорами з кількістю обчислювальних ядер більше двох для продуктивної роботи алгоритмів ущільнення було б цілком справедливо виконувати алгоритми ущільнення за допомогою розпаралелювання.[5]

Оскільки мова йде про ущільнення текстової інформації необхідно проводити компресію без втрат, в іншому випадку результуючий файл буде неможливо повернути до виду вхідного, і інформація буде спотворена при декомпресії, що є неприпустимим.[2]

Найбільш вдалими для виконання поставленої мети з урахуванням специфіки оброблюваних даних будуть статичні методи алгоритмів ущільнення, що кодують символи по заданим ймовірностям появи комбінації в тексті з розподіленням ймовірностей або ж алгоритми, що базуються на словниковому ущільненні[3], що використовують шматочки закодованих або відомих декодеру даних.

До статичних методів відносять алгоритм Хаффмана, код Шеннона-Фано та арифметичне кодування. Розглянемо статичні методи на прикладі алгоритму Хаффмана що відображає основну ідею ущільнення без втрат.

Припустимо що нам відомі ймовірності p_i появи кожного символу вхідного алфавіту a_i.

Оберем два символа з найменшою ймовірністю і замінемо його на новий символ $\{a_i, a_j\}$ ймовірність появи якого дорівнює сумі ймовірностей цих двох символів. Кожна пара замінених символів буде представлятись як 0 та 1. Якщо це символ виду $\{a_i, a_j\}$, то відповідний код додамо до обох символів. Таким чином в результаті кожному символу буде відповідати код довжина якого залежить від того, наскільки часто він використовується.[4] Принцип такий: чим

частіше символ зустрічається, тим коротшим буде його код, відповідно символ, що зустрічається у повідомленні найрідше буде мати найдовшу кодову комбінацію. Схожий принцип має адаптивний метод Хаффмана в основі якого «дерево Хаффмана», у якому кожному символу визначається вага, яка визначається відношенням ваги символу до суми всіх ваг. Метод може ефективно працювати при розпаралелюванні якщо кожен потік буде опрацьовувати визначені алгоритмом частини повідомлення будуючи спільне дерево, яке в свою чергу буде адаптуватись динамічно змінюючи ваги символів, що зустрічаються, що у результаті пришвидшить пошук оптимального кодування символів для ущільнення вихідного повідомлення.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Claude E. Shannon. The Mathematical Theory of Communication / Claude E. Shannon, Weaver Warren. – University of Illinois Press, Urbana, 1963. – 63с.
2. Ватолин Д., Ратушняк А., Смирнов М, Юкин В. Методы сжатия данных. Устройство архиваторов, сжатие изображений и видео. Диалог-МИФИ, 2003. 384 стр.
3. Кнут Д. Искусство программирования, том 2. Получисленные алгоритмы. / Д. Кнут. – Изд. 3-е; пер. с англ. – М. : Издательский дом «Вильямс», 2007. – 832 с.
4. Блейхут Р. Теория и практика кодов, контролирующих ошибки / Р. Блейхут ; пер. с англ. – М. : Мир, 1986. – 576 с.
5. Семеренко В. П. Темпоральні моделі паралельних обчислень / В. П. Семеренко // Austrian Journal of Technical and Natural Sciences. – January-February, 2014. – № 1. – P. 13–25.

Дмитро Олегович Пущал – студент групи 1КІ-16мс, факультет інформаційних технологій та комп'ютерної інженерії, Вінницький національний технічний університет, Вінниця, e-mail hancer1996@gmail.com

Dmyrto O. Pushchal – student, Department of computer technique, Vinnytsia National Technical University, Vinnytsia, e-mail: hancer1996@gmail.com