УДК 004.42

А.М. ГЛИБОВЕЦЬ
Національний університет «Києво-Могилянська академія», Київ, Україна,
e-mail: *a.glybovets@ukma.edu.ua*.

В.О. ДІДЕНКО
Національний університет «Києво-Могилянська академія», Київ, Україна,
e-mail: *verochka1998@gmail.com*.

ПОБУДОВА УЗАГАЛЬНЕНИХ СУФІКСНИХ ДЕРЕВ НА РОЗПОДІЛЕНИХ ПАРАЛЕЛЬНИХ ПЛАТФОРМАХ

Анотація. Запропоновано алгоритм побудови узагальнених суфіксних дерев з використанням розподілених паралельних платформ, який є оптимальним з погляду як часової складності, так і використання пам'яті. Розподілений підхід до побудови дає змогу працювати з великими алфавітами та дуже довгими рядками. Алгоритм є ефективним щодо масштабованості на розподілених паралельних платформах і підтримує суфікси індексування для різноманітних довгих рядків, починаючи від одного довгого рядка до кількох довгих рядків різної довжини.

Ключові слова: суфіксне дерево, узагальнене суфіксне дерево, ERa, алгоритм, паралельна побудова суфіксного дерева.

ВСТУП

Суфіксне дерево [1, 2] є фундаментальною структурою даних для роботи із символьною інформацією і використовується під час розв'язування багатьох задач оптимізації оброблення тексту великого розміру. Суфіксне дерево (СД) зберігає рядок шляхом індексації всіх можливих суфіксів цього рядка та застосовується у різноманітних системах для забезпечення швидких операцій над рядками (зіставлення фраз, пошук найдовшого повторюваного підрядка у послідовності символів або виявлення максимальної кількості повторюваних фраз у довгій послідовності символів).

З огляду на інтенсивне зростання обсягів даних, які потребують машинного оброблення, дослідженню алгоритмів побудови оптимального СД приділяють велику увагу [3–6]. Значних зусиль докладено до побудови паралельних алгоритмів СД на основі MPI [3]. Однак, вони мають обмеження з погляду масштабованості та відмовостійкості у тому разі, коли здійснюється введення великих за розміром даних. Водночас постійно зростає попит на ефективні алгоритми для побудови суфіксних дерев на розподілених паралельних платформах, як-от Hadoop [7] і Apache Spark [8].

У цій роботі представлено ефективний і високомасштабований алгоритм побудови узагальнених дерев суфіксів на розподілених паралельних платформах. Він складається з двох основних етапів: поділу паралельного піддерева та побудови паралельного піддерева.

На першому етапі реалізовано нову стратегію розповсюдження даних. Далі запропоновано ефективний алгоритм поділу піддерева, який будує LCP-масив [9–12] у паралельний спосіб, тобто паралельно обчислює, скільки разів зустрічається підрядок. Щоб покращити балансування навантаження та зменшити витрати на читання даних, передбачено ефективну стратегію розподілу завдань між паралельними процесами.

На другому етапі під час побудови паралельного піддерева використано структуру даних LCP-Range та багатосторонній алгоритм сортування LCP-Merge [9–12] для паралельної побудови LCP-масиву. Також модифіковано