

ВИКОРИСТАННЯ СУФІКСНИХ ДЕРЕВ ДЛЯ ПОКРАЩЕННЯ ОПЕРАЦІЙ ПОШУКУ В СИСТЕМАХ УПРАВЛІННЯ БАЗ ДАНИХ

Д.В. ЗВАЖІЙ

АНОТАЦІЯ. Сучасні системи управління базами даних (СУБД) вимагають високої швидкості операцій пошуку для ефективного виконання бізнес-задач. Традиційні підходи індексації, такі як В-дерева та хеш-таблиці, можуть бути недостатньо ефективними для певного набору запитів. У цій доповіді буде розглянуто альтернативний підхід із використанням суфіксних дерев, які можуть значно покращити швидкість пошукових операцій[1].

Означення. *Суфіксне дерево T стрічки S довжиною t це кореневе орієнтоване дерево, що має t листків пронумерованих $1..t$. Кожна внутрішня вершина, окрім кореня, має щонайменше два ребра, кожне ребро марковане непорожнім підрядком S . Жодна вершина не може мати ребра, що починається з одного й того самого символу. Головною особливістю суфіксного дерева є те, що конкатенація маркувань на ребрах, починаючи від кореня до листка i , генерує суфікс стрічки S , починаючи з i -го символу.[2]*

Для задачі індексації потрібно працювати з колекціями стрічок, тому в таких випадках краще використовувати узагальнені суфіксні дерева. Для побудови узагальнених суфіксних дерев також використовують термінальний символ, зазвичай “\$”. Унікальність термінального символу дозволяє уникнути ситуацій, коли суфікс стає префіксом іншого суфіксу.

Також потрібно ввести набір операцій, час виконання яких потрібно покращити.

Означення 1. *Пошук стрічки (повне співпадіння) - результатом пошуку стрічки S в колекції C є всі стрічки $S_1, S_2, \dots, S_n$, кожен символ яких співпадає з кожним символом S , $S_n[0] = S[0], S_n[1] = S[1], S_n[t-1] = S[t-1]$*

Означення 2. *Пошук підстрічок - результатом пошуку стрічки S в колекції C є всі стрічки $S_1, S_2, \dots, S_n$, для кожної з яких існує таке i , що $S_n[i] = S[0], S_n[i+1] = S[1], S_n[i+t-1] = S[t-1]$.*

Означення 3. Підстрічка не входить у стрічку - результатом пошуку стрічки S в колекції C є всі стрічки $S_1, S_2, \dots, S_n$, для кожної з яких не існує такого i , що $S_n[i] = S[0], S_n[i+1] = S[1], S_n[i+m-1] = S[m-1]$.

Означення 4. Починається з - результатом пошуку стрічки S в колекції C є всі стрічки $S_1, S_2, \dots, S_n$, для кожної з яких виконується наступне $S_n[0] = S[0], S_n[1] = S[1], S_n[m-1] = S[m-1]$.

Означення 5. Не починається з - результатом пошуку стрічки S в колекції C є всі стрічки $S_1, S_2, \dots, S_n$, для кожної з яких виконується наступне

$$S_n[0] \neq S[0], S_n[1] \neq S[1], S_n[m-1] \neq S[m-1]$$

Означення 6. Закінчується на - результатом пошуку стрічки S в колекції C є всі стрічки $S_1, S_2, \dots, S_n$, для кожної з яких виконується наступне $S_n[i] = S[0], S_n[i+1] = S[1], S_n[i+m-1] = S[m-1]$, де n довжина довільної стрічки S_n та виконується $i+m=n$

Означення 7. Не закінчується на - результатом пошуку стрічки S в колекції C є всі стрічки $S_1, S_2, \dots, S_n$, для кожної з яких не виконується наступне $S_n[i] = S[0], S_n[i+1] = S[1], S_n[i+m-1] = S[m-1]$, де n довжина довільної стрічки S_n та виконується $i+m=n$

Оцінки складності для описаних вище функцій виглядатимуть наступним чином[2]:

Назва операції	Оцінка складності операції (теоретична)
Пошук стрічки	$O(m)$
Пошук підстрічок	$O(m)$
Підстрічка не входить у стрічку	$O(m)$
Починається з	$O(m)$
Не починається з	$O(m)$
Закінчується на	$O(m)$
Не закінчується на	$O(m)$

де m - довжина стрічки, що приходить на вхід функції пошуку.

Гіпотеза. Впровадження суфіксних дерев у якості ідексів для СУБД зможе знизити час, який необхідний для виконання пошукових запитів такі як: “Пошук стрічки”, “Пошук підстрічок”, “Підстрічка не входить у стрічку”, “Починається з”, “Не починається з”, “Закінчується на”, “Не закінчується на”.

ЛІТЕРАТУРА

- [1] Haoyu C., Ge N. Yidong H., Xuedong H., Wentao X. *Full-text search engine with suffix index for massive heterogeneous data* // Information Systems. — 2022. — С. 104.

- [2] Gusfield D. *Algorithms on strings, trees, and sequences: computer science and computational biology*. — Cambridge: Cambridge University Press, 1997. — 96 с, 123 с.

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ “КИЄВО-МОГИЛЯНСЬКА АКАДЕМІЯ”, Київ, Україна
Email address: `d.zvazhii@ukma.edu.ua`