ДОДАТОК А

Перелік джерел посилання за науковими напрямами керівника та науковців кафедри програмної інженерії

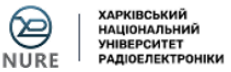
10. Falatiuk H., Shirokopetleva M., Dudar Z. Investigation of Architecture and Technology Stack for e-Archive System. 2019 IEEE International Scientific-Practical Conference Problems of Infocommunications, Science and Technology (PIC S&T), Kyiv,Ukraine,8-11October2019.2019.,URL: https://doi.org/10.1109/picst47496.2019.9061407 (дата звернення: 05.05.2024).

12. DATA EXCHANGE MODEL IN THE INTERNET OF THINGS CONCEPT / I. Afanasieva et al. Telecommunications and Radio Engineering. 2019. Vol. 78,no. 10.P. 869–878.URL: https://doi.org/10.1615/telecomradeng.v78.i10.30 (date of access: 05.05.2024).

13. Аналіз результатів бенчмаркіунгу
URL: https://doi.org/10.1615/telecomradeng.v78.i10.30 (date of access: 05.05.2024)

ДОДАТОК Б

Слайди презентації



Дослідження методів та алгоритмів роботи з повнотекстовими індексами

Сорокін Володимир Віталійович., ІПЗм-22-5

Науковий керівник: проф. Смеляков К. С

14 червня 2024

Дослідження

Актуальність та стан розвитку галузі: Дослідження в області повнотекстового індексування є особливо актуальним, оскільки воно забезпечує основу для ефективного пошуку в великих базах даних і системах управління контентом. Традиційні методи не можуть ефективно обробляти такі запити через велику кількість даних і їх неструктурований характер. Така ситуація підкреслює необхідність розробки і вдосконалення технологій повнотекстового індексування.

Чітке визначення напряму дослідження: Основна мета полягає у визначенні методів, які забезпечують оптимальне співвідношення швидкості виконання запитів та точності результатів, при мінімальному використанні системних ресурсів.

Об'єкт дослідження: Методи та технології повнотекстового індексування, їх застосування у системах управління базами даних і пошукових системах.

Постановка задачі

- Ця робота спрямована створення дорожньої карти для розробників що до знаходження найефективнішого матоду пошуку за повнотекстовими індексами, також мета роботи – модернізувати деякі підходи для досягнення кращого результату у пошуку.
- Реалізація наукового дослідження складається з наступних етапів:

– аналіз предметної області;
– аналіз методів пошуку за повнотекстовими індексами;
– розгляд особливостей кожного методу;
– порівняльний аналіз та виведення критеріїв порівняння;
– модифікація пошуку за файловими системами;
– розробка бенчмарку для методів;
– розробка розбиття на підфайли;
– формування висновків.

Методологія

Опис використаних методів дослідження: Задопомогою програмного застосунку FullTextIndexG, було виміряно середній час у мікросекундах, середне відхилення та кількіст «збірок сміття».

Інструментарій та технології, використані в роботі:

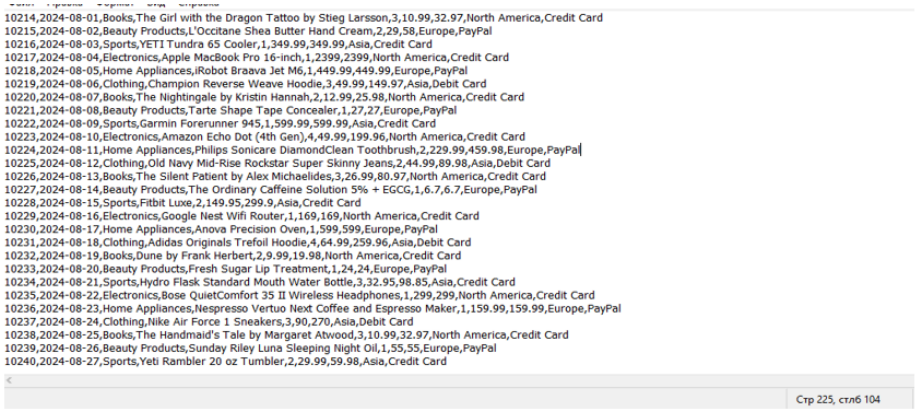
- Платформи: .Net
- Засоби тестування: FullTextIndexG, BenchmarkDotNet.
- Мова програмування C#

Дослідження проблематики

- Було досліджено існуючі методи пошуку у повнотекстових індексах
- Досліджено приклади реалізації простого лінійного пошуку та інших алгоритмів
- Виявлено неефективність простого лінійного пошуку на великих наборах даних
- Виявлено необхідність у тестуванні продуктивності різних методів індексування та пошуку на великих обсягах даних

Тестовий дата-сет

Отже у нас за основу взято приклад дата-сету на 10 000 слів.



```
10214,2024-08-01,Books,The Girl with the Dragon Tattoo by Stieg Larsson,3,10.99,32.97,North America,Credit Card
10215,2024-08-02,Beauty Products,L'Occitane Shea Butter Hand Cream,2,29,58,Europe,PayPal
10216,2024-08-03,Sports,YETI Tundra 65 Cooler,1,349.99,349.99,Asia,Credit Card
10217,2024-08-04,Electronics,Apple MacBook Pro 16-inch,1,2399,2399,North America,Credit Card
10218,2024-08-05,Home Appliances,iRobot Braava Jet M6,1,449.99,449.99,Europe,PayPal
10219,2024-08-06,Clothing,Champion Reverse Weave Hoodie,3,49.99,149.97,Asia,Debit Card
10220,2024-08-07,Books,The Nightingale by Kristin Hannah,2,12.99,25.98,North America,Credit Card
10221,2024-08-08,Beauty Products,Tarte Shape Tape Concealer,1,27,27,Europe,PayPal
10222,2024-08-09,Sports,Garmin Forerunner 945,1,599.99,599.99,Asia,Credit Card
10223,2024-08-10,Electronics,Amazon Echo Dot (4th Gen),4,49.99,199.96,North America,Credit Card
10224,2024-08-11,Home Appliances,Philips Sonicare DiamondClean Toothbrush,2,229.99,459.98,Europe,PayPal
10225,2024-08-12,Clothing,Old Navy Mid-Rise Rockstar Super Skinny Jeans,2,44.99,89.98,Asia,Debit Card
10226,2024-08-13,Books,The Silent Patient by Alex Michaelides,3,26.99,80.97,North America,Credit Card
10227,2024-08-14,Beauty Products,The Ordinary Caffeine Solution 5% + EGCG,1,6.7,6.7,Europe,PayPal
10228,2024-08-15,Sports,Fitbit Luxe,2,149.95,299.9,Asia,Credit Card
10229,2024-08-16,Electronics,Google Nest Wifi Router,1,169,169,North America,Credit Card
10230,2024-08-17,Home Appliances,Anova Precision Oven,1,599,599,Europe,PayPal
10231,2024-08-18,Clothing,Adidas Originals Trefoil Hoodie,4,64.99,259.96,Asia,Debit Card
10232,2024-08-19,Books,Dune by Frank Herbert,2,9.99,19.98,North America,Credit Card
10233,2024-08-20,Beauty Products,Fresh Sugar Lip Treatment,1,24,24,Europe,PayPal
10234,2024-08-21,Sports,Hydro Flask Standard Mouth Water Bottle,3,32.95,98.85,Asia,Credit Card
10235,2024-08-22,Electronics,Bose QuietComfort 35 II Wireless Headphones,1,299,299,North America,Credit Card
10236,2024-08-23,Home Appliances,Nespresso Vertuo Next Coffee and Espresso Maker,1,159.99,159.99,Europe,PayPal
10237,2024-08-24,Clothing,Nike Air Force 1 Sneakers,3,90,270,Asia,Debit Card
10238,2024-08-25,Books,The Handmaid's Tale by Margaret Atwood,3,10.99,32.97,North America,Credit Card
10239,2024-08-26,Beauty Products,Sunday Riley Luna Sleeping Night Oil,1,55,55,Europe,PayPal
10240,2024-08-27,Sports,Yeti Rambler 20 oz Tumbler,2,29.99,59.98,Asia,Credit Card
```

Стр 225, стлб 104

Це гарний набір даних адже він не величезний як наприклад бібліотеки або не доволі малі як збірка відгуків, для більшої зручності та швидкості, дані описую збірку американських навісних каналів

Simple search

- Лінійний перебір потребує значних обчислювальних ресурсів, особливо при великій кількості запитів, що може призвести до перевантаження системи.

Method	Mean	Error	StdDev	Allocated
SimpleSearch	4.269 s	0.5880 s	0.0910 s	1 MB

Методи пошуку

Клас FragmentBasedIndex використовує підхід, який дещо відрізняється від традиційних методів індексування, де індексуються окремі слова. Замість цього, цей метод зосереджується на індексації фрагментів тексту. Це може бути корисно для оптимізації пошуку в документах, які містять багато тексту, або коли потрібно зберігати більше контексту навколо ключових слів.

Клас FullTextIndexWithPositions, який ми використовуємо в своєму проекті, є реалізацією повнотекстового індексу, який, на відміну від простіших індексів, зберігає не тільки інформацію про наявність токена (слова) в документі, але й точні позиції цих токенів у тексті документа.

Аналіз результатів

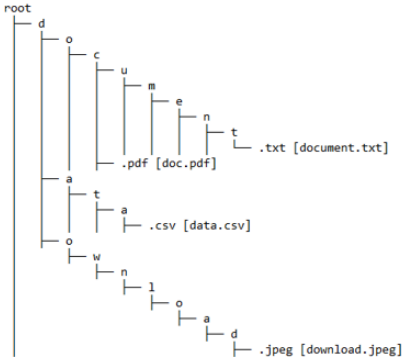
- На зображенні відображено таблицю результатів бенчмаркінгу для чотирьох різних методів індексування документів, а також читання датасету. Результати включають виміри таких параметрів, як середній час виконання (Mean), похибка (Error), стандартне відхилення (StdDev), співвідношення до базового методу (Ratio), варіація співвідношення (RatioSD), кількість використань пам'яті у поколінні Gen 0 і загальний обсяг виділеної пам'яті (Allocated).

```
InvocationCount=5  IterationCount=5  UnrollFactor=1
WarmupCount=1

|          Method |     Mean |      Error |   StdDev | Ratio | RatioSD |    Gen 0 | Allocated |
|---------------- |---------:|-----------:|---------:|------:|--------:|---------:|----------:|
|     ReadDataset | 904.4 us | 127.46 us | 33.10 us |  1.00 |    0.00 |        - |    556 KB |
|     SimpleIndex | 390.6 us |  20.57 us |  5.34 us |  0.43 |    0.02 |        - |    419 KB |
| PositionedIndex | 824.5 us | 224.51 us | 58.30 us |  0.91 |    0.07 |        - |    662 KB |
|   FragmentIndex | 684.9 us | 322.34 us | 49.88 us |  0.75 |    0.08 | 200.0000 |  1,303 KB |
```

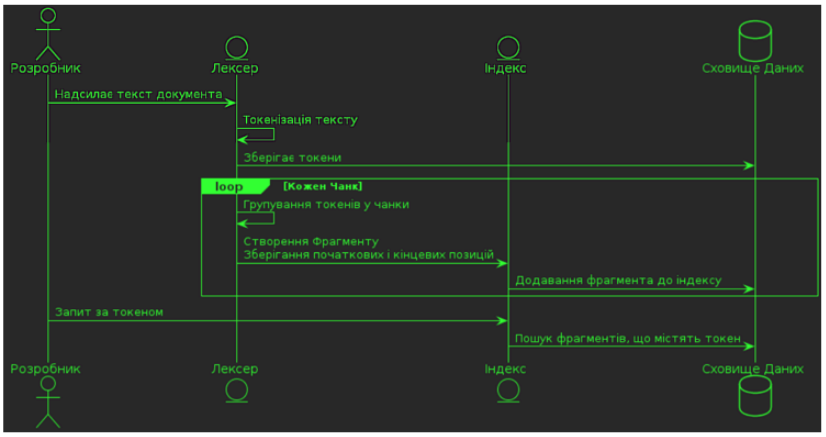
Модифікація файлового пошуку

- Метод використання префіксного дерева (trie) для модифікації сховища повнотекстових індексів пропонує елегантне рішення цієї проблеми. Префіксні дерева дозволяють швидко знаходити слова або фрази за їхніми початковими символами, що є ідеальним для пошуку у великих текстових базах даних.
- Розглянемо більш детально процес імплементації та роботи методу.



Метод фрагментації тексту

Фрагментація тексту дозволяє системам швидше знаходити необхідні дані, оскільки пошук ведеться лише в окремих, менших фрагментах тексту, а не в усьому документі або наборі документів. Це різко зменшує обсяг даних, які потрібно просканувати під час пошукового запиту, тим самим підвищуючи швидкість обробки.



Аналіз часу обробки

- SimpleSearch виявився найповільнішим у всіх трьох тестових сценаріях, що підтверджує його обмежену ефективність для швидкого доступу до даних.
- FragmentsSearch і WithPositionsIndexedSearch демонструють вражаючу швидкість для простих запитів, що свідчить про високу оптимізацію індексації та пошуку у фрагментах даних.
- FullTextIndexedSearch є ефективним для менш складних запитів, але його продуктивність знижується при більш складних запитах, як показано в тесті "where".
- WithPositionsIndexedSearch забезпечує найкращі результати для складного запиту "where", що підкреслює його здатність ефективно обробляти високу складність запитів з мінімальним часом відгуку.

Method	Query	Mean	Error	StdDev	Ratio	Gen 0	Allocated
SimpleSearch	Books	1,830,737.89 ns	815,112.256 ns	211,682.064 ns	1.000	15.6250	69,265 B
FragmentsSearch	Books	89.08 ns	15.276 ns	3.967 ns	0.000	0.0421	176 B
FullTextIndexedSearch	Books	98.54 ns	1.307 ns	0.202 ns	0.000	0.0497	208 B
WithPositionsIndexedSearch	Books	99.36 ns	1.501 ns	0.232 ns	0.000	0.0592	248 B
SimpleSearch	assessment	1,914,779.69 ns	28,848.025 ns	4,464.264 ns	1.000	15.6250	69,265 B
FragmentsSearch	assessment	88.80 ns	9.387 ns	1.453 ns	0.000	0.0421	176 B
FullTextIndexedSearch	assessment	87.05 ns	4.832 ns	1.255 ns	0.000	0.0421	176 B
WithPositionsIndexedSearch	assessment	88.88 ns	6.201 ns	1.610 ns	0.000	0.0516	216 B
SimpleSearch	where	1,623,642.62 ns	14,599.704 ns	3,791.497 ns	1.000	15.6250	69,265 B
FragmentsSearch	where	6,867.84 ns	150.968 ns	39.206 ns	0.004	0.1068	464 B
FullTextIndexedSearch	where	8,237.68 ns	302.196 ns	78.479 ns	0.005	0.0916	400 B
WithPositionsIndexedSearch	where	272.12 ns	12.966 ns	3.367 ns	0.000	0.1259	528 B

Висновки

- Ефективність методів: WithPositionsIndexedSearch виявився найефективнішим для складних запитів, підтримуючи високу швидкість обробки даних із мінімальним часом відгуку.
- Оптимізація ресурсів: FragmentsSearch та FullTextIndexedSearch продемонстрували високу продуктивність для менш складних запитів із раціональним використанням системних ресурсів.
- Складність операцій: SimpleSearch показав найгірші результати з великим часом відгуку, що робить його менш придатним для вимогливих операцій.
- Використання пам'яті: Індексація з позиціонуванням потребувала більшого обсягу пам'яті, але забезпечила кращу продуктивність для великої кількості запитів.

ДОДАТОК В

Звіт результатів перевірки на унікальність тексту в базі ХНУРЕ

Ім'я користувача:
Кардаш Євген Вікторович каф.ПІ

ID перевірки:
1016357983

Дата перевірки:
13.06.2024 21:48:05 EEST

Тип перевірки:
Doc vs Internet + Library

Дата звіту:
14.06.2024 06:13:26 EEST

ID користувача:
100013622

Назва документа: 2024_М_ПІ_ІПЗм_22_5_Сорокін_В_В_скорочений

Кількість сторінок: 43 **Кількість слів:** 7595 **Кількість символів:** 58528 **Розмір файлу:** 743.55 KB **ID файлу:** 1016162429

25.9%
Схожість

Найбільша схожість: 16.8% з джерелом з Бібліотеки (ID файлу: 1016096962)

24.9% Джерела з Інтернету	121	Сторінка 45
16.9% Джерела з Бібліотеки	6	Сторінка 45

0% Цитат

Вилучення цитат вимкнене

Вилучення списку бібліографічних посилань вимкнене

0%
Вилучень

Немає вилучених джерел

ДОДАТОК Г

Апробація результатів роботи

УДК 004.415:004.2

ДОСЛІДЖЕННЯ МЕТОДІВ ТА АЛГОРИТМІВ РОБОТИ З ПОВНОТЕКСТОВИМИ ІНДЕКСАМИ

Сорокін В. В.
Науковий керівник – проф.
Смеляков К. С.
Харківський національний університет радіоелектроніки, каф. ПІ
м. Харків, Україна

У сучасному світі дані відіграють дуже важливу роль в нашому житті, вони надають нам різноманітну інформацію різних напрямків і форм, тому обсяг даних зростає і взаємодіяти з ними стає все складніше. Для ефективної взаємодії з великими обсягами даних розробники використовують індексацію, вона застосовується для всіх типів даних, і повнотекстові індекси є найбільш поширеними, але вони мають великий обсяг і складну структуру.

Однак, незважаючи на всі переваги, які надають повнотекстові індекси, існує ряд складнощів і питань, особливо коли мова заходить про вибір оптимального методу пошуку. Проблема вибору ефективного методу пошуку за повнотекстовими індексами є актуальною з кількох причин. По-перше, із зростанням обсягу цифрових даних, особливо текстової інформації, що генерується користувачами в інтернеті, соціальних мережах, в наукових базах даних, архівах та корпоративних системах, виникає величезна потреба у швидкому та точному доступі до цієї інформації. Повнотекстові індекси дозволяють організувати великі набори даних таким чином, щоб користувачі могли ефективно здійснювати пошук, знаходячи не тільки документи, що містять певні слова або фрази, але й оцінюючи релевантність цих документів до запиту.

Крім того, збільшення обсягу даних створює виклики, пов'язані з обробкою запитів та видачою результатів пошуку. Традиційні методи, такі як простий лінійний пошук або навіть складніші алгоритми, можуть бути неефективними або надмірно ресурс затратними при роботі з великими

дата-сетами. Це призводить до необхідності вибору методу пошуку, який би оптимізував процес індексації та пошуку, забезпечуючи при цьому високу точність та швидкість обробки.

Нарешті, з урахуванням різноманітності запитів користувачів та складності інформаційних потреб, методи пошуку повинні бути гнучкими та вміти адаптуватися до різних типів даних та запитів. Це означає, що методи пошуку мають враховувати не тільки ключові слова, але й контекст, в якому ці слова використовуються, а також здатні обробляти складні запити, які включають логічні оператори або регулярні вирази.

Аналізуючи предметну область по використанню та роботі із повнотекстовими індексами – можна зрозуміти що зазвичай це є певні збірки, будь то музики, текстових матеріалів таких як книги чи статті, також повнотекстові індекси часто використовуються для аналізу новин та навісних статей яких може буде доволі багато у вибірці, отже виходячи з цього можна зрозуміти що варто розглядати методи пошуку які в першу чергу зосереджені на роботі із великими обсягами даних.

Виходячи із цього надалі було розглянуто та порівняно методи у контексті ефективності роботи із великими обсягами даних. У висновку було з’ясовано що краще за все працює файловий пошук та його модифікована версія, модифікація була проведена наступним чином - припустимо в нас великий об’єм даних з яким не можливо працювати через оперативну пам’ять, тоді ми будемо використовувати метод розміщення індексів у файлову систему на диску, тож тепер в нас постає проблема пошуку об’єкту по файловій системі.

Отже, однією з ключових проблем є забезпечення швидкого і точного доступу до інформації. Існуючі методи повнотекстового індексування як вже з’ясувано є неефективними при роботі з великими обсягами даних, оскільки вони часто вимагають значного часу для пошуку та фільтрації інформації.

Цю проблему можна вирішити лінійним пошуком або бінарним та навіть у випадку використання хеш таблиць – пошук буде лінійно сповільнюватись відносно зростаючого об’єму файлів, тож треба розробити більш ефективний

метод для роботи із пошуком у файловій системі.

Метод використання префіксного дерева (trie) для модифікації сховища повнотекстових індексів пропонує елегантне рішення цієї проблеми. Префіксні дерева дозволяють швидко знаходити слова або фрази за їхніми початковими символами, що є ідеальним для пошуку у великих текстових базах даних.

Розглянемо більш детально процес імплементації та роботи методу.

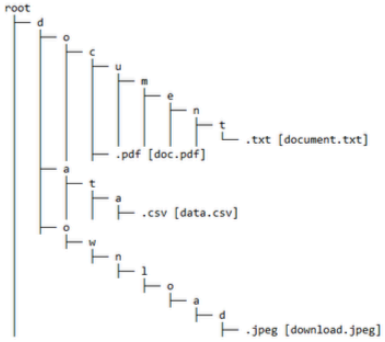
Спочатку створюється префіксне дерево, де кожен вузол представляє окремий символ у слові або фразі. Вузли пов'язані таким чином, що кожен шлях від кореня до вузла відображає унікальне слово або частину слова.

Це дерево інтегрується із системою сховища даних, що дозволяє індексувати великі обсяги текстової інформації.

Кожен документ або запис у повнотекстовому індексі проходить через процес індексації, де його текст розбивається на слова або фрази, які потім додаються у префіксне дерево. Індекси або підіндекси зберігаються в вузлах дерева для швидкого доступу до відповідних документів.

Інтеграція префіксного дерева з сховищем повнотекстових індексів пропонує значне покращення в швидкості та ефективності пошуку індексованих файлів, особливо при роботі з великими обсягами даних. Це рішення не тільки оптимізує пошук та індексацію даних, але й забезпечує покращення швидкості знаходження інформації через обхід у н-етерацій.

На рисунку 1 зображене схематичний вигляд файлової системи після створення її за нашим методом.



Рисунку 1 –Схематичний вигляд файлової системи

Підсумовуючи, метод використання префіксного дерева для модифікації сховища повнотекстових індексів є незвичним але високоефективним підходом, який пропонує ряд переваг у порівнянні з іншими методами пошуку в файлових системах.

Завершуючи аналіз, було зосереджено увагу на методі використання префіксних дерев для модифікації файлових систем. Цей метод виявився надзвичайно ефективним у покращенні швидкості пошуку індексованих файлів, особливо у великих обсягах даних.

Загалом щоб остаточно визначити найефективніший метод пошуку – та більш детально дослідити метод префіксних дерев – необхідно буде провести додаткові дослідження.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ:

1. Aniruddha Karajgi. Understanding Prefix Trees. 2022.: https://polaris000.medium.com/understanding-prefix-trees-13da74b3cafb (дата звернення: 16.05.2023).

ДОДАТОК Д

Експертний висновок результатів перевірки кваліфікаційної роботи на відповідність оформлення вимогам ДСТУ 3008:2015

Експертний висновок результатів перевірки кваліфікаційної роботи

студент (посада) | програмної інженерії (кафедра) | ІПЗм-22-5 (група)

Сорокін Володимир Віталійович
(прізвище, ім'я, по батькові)

Зауваження

Пункт ДСТУ 3008-2015	Зміст пункту	Сторінка кваліфікаційної роботи
1	2	3
	7.1 Загальні положення	
	7.3 Нумерація сторінок звіту	
	7.4 Нумерація розділів, підрозділів, пунктів, підпунктів	
	7.5 Рисунки	
	7.6 Таблиці	
	7.7 Переліки	
	7.8 Примітки	
	7.9 Виноски	
	7.10 Формули та рівняння	
	7.11 Посилання	
	7.13 Список авторів	
	7.14 Скорочення та умовні познаки	
	7.15 Додатки	

зауважень немає

Експерт _____ (підпис) Олена ОЛІЙНИК (прізвище, ініціали)

15.06.2024