Struktury Baz Danych

Projekt 2: Implementacja organizacji pliku

Rafał Kajomof 193322

23 Grudnia 2024

1. Cel projektu

Celem projektu jest implementacja wybranej indeksowej organizacji pliku, aby trwale zapisać dane oraz wydajnie nimi manipulować, i przeprowadzenie eksperymentów mających na celu ustalenie wpływu parametrów implementacyjnych na złożoność poszczególnych operacji w pliku, takich jak wstawianie rekordu, aktualizowanie rekordu, czy wyświetlenie rekordu. Program symuluje odczyty i zapisy blokowe między pamięcią operacyjną i masową, aby pokazać jak zachowa się w przypadku pracy z dużymi zbiorami danych w rzeczywistych warunkach. Wyniki eksperymentów zostaną przedstawione w formie wykresów i tabeli oraz poddane analizie.

2. Opis zastosowanej metody

W tym projekcie zaimplementowano organizację indeksowo-sekwencyjną do organizacji pliku z danymi. Jest to struktura, która pozwala na wyszukiwanie za pomocą indeksu oraz sekwencyjnie przechowuje dane w części głównej. Klucze do rekordów wraz z indeksami stron są przechowywane w osobnej strukturze indeksowej, która umożliwia szybkie wyszukiwanie pozycji rekordu w pliku. Dane rekordów są przechowywane w postaci sekwencyjnej w części głównej.

3. Implementacja

Projekt został zaimplementowany w całości w języku C++. Program przechowuje dane w postaci trzech plików w formacie binarnym: pliku indeksowym, pliku z danymi rekordów w części głównej i pliku z danymi rekordów w części nadmiarowej. Program uporządkowuje rekordy rosnąco. Do reorganizacji pliku wykorzystano algorytm sortujący metodą scalania naturalnego, dopasowany do potrzeb metody indeksowo-sekwencyjnej.

Program reorganizuje plik, gdy część nadmiarowa przekroczy pewien ustalony limit, i możliwa jest także reorganizacja pliku na żądanie. Zaimplementowano operacje wstawiania rekordu, pokazywania rekordu, aktualizowania rekordu, pokazania całej struktury pliku i reorganizacji pliku.

Każda strona dyskowa ma wielkość 4 rekordów.

4. Specyfikacja formatów plików

Każdy plik ma postać binarną. W pliku indeksowym, każda pozycja odpowiada jednej stronie w pliku z częścią główną i ma postać dwóch ośmiobajtowych liczb, z których pierwsza jest kluczem pierwszego rekordu występującego na danej stronie, a druga jest indeksem strony, na której się znajduje.

W pliku z częścią główną i nadmiarową znajduje się binarna reprezentacja rekordów zawierająca klucz rekordu, ilość liczb w rekordzie, liczby należące do rekordu, oraz wskaźnik na kolejny rekord w części nadmiarowej, jeśli taki rekord istnieje.

Plik testowy ma postać tekstową. Składa się on z komend, takich jak:

- insert,
- show,

- list,
- update,
- status,
- sort.

5. Opis sposobu prezentacji wyników działania programu

Program oczekuje komend na standardowym strumieniu wejściowym. Po wprowadzeniu każdej komendy, program wypisuje na standardowym strumieniu wyjściowym informacje o wyniku komendy i operacjach na dysku przeprowadzonych podczas wykonywania operacji związanych z komendą. Dodatkowo, program może wyświetlić szczegółowe informacje o operacjach na dysku.

6. Opis eksperymentu

Eksperyment miał na celu zbadanie zależności złożoności operacji wykonywanych na pliku od poszczególnych paramentrów implementacyjnych algorytmu. W eksperymentach przeprowadzono następujące kroki:

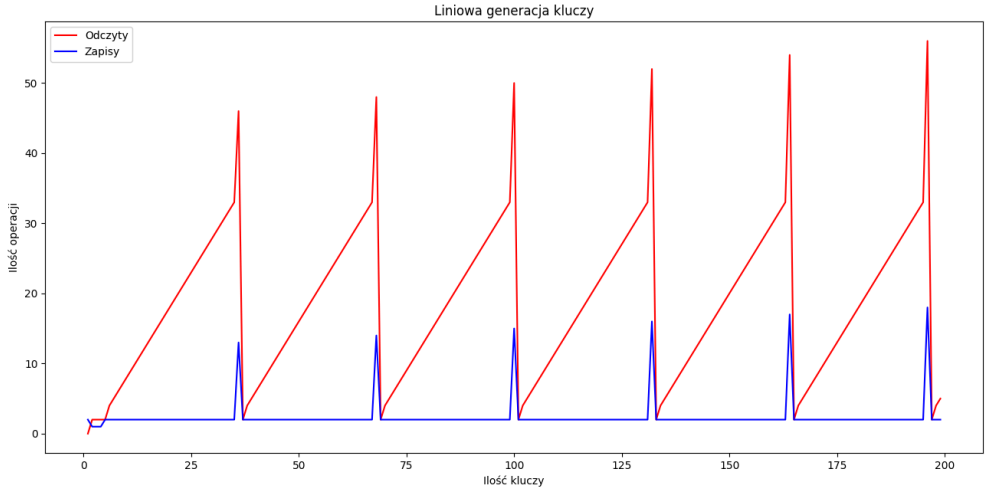
1. Generowanie danych
2. Wstawianie danych
3. Zebranie danych

Dla różnych limitów wielkości pliku części nadmiarowej wygenerowano rekordy w sposób liniowy i losowy, i wstawiono je po kolei, reorganizując plik na podstawie wielkości pliku części nadmiarowej, mierząc liczbę odczytów i zapisów stron dyskowych. Tak zebrane dane dla różnych limitów wielkości części nadmiarowej porównano.

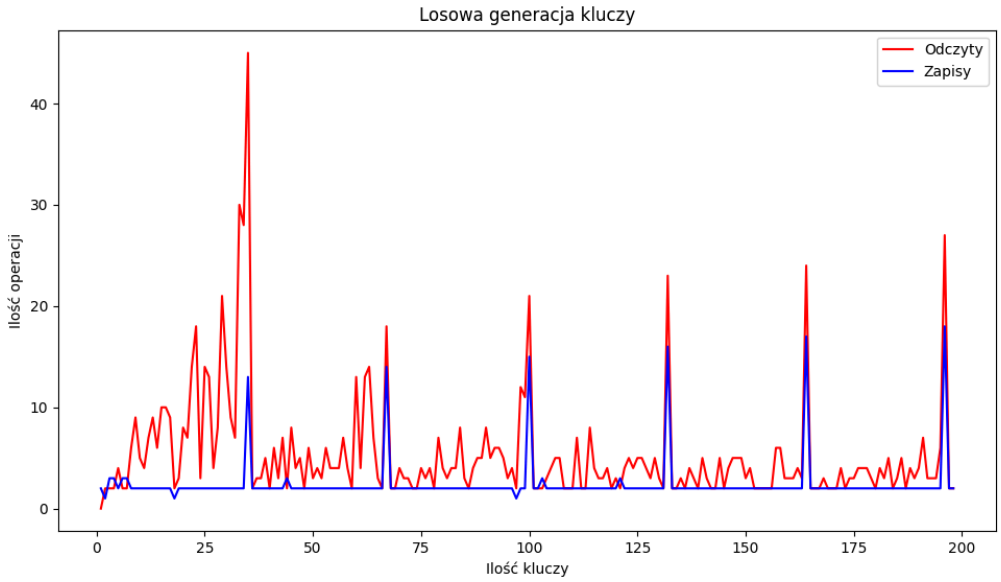
Wyniki eksperymentu zostały przedstawione w formie tabeli oraz na wykresie.

Numer rekordu	Liczba odczytów	Liczba zapisów
1	0	2
2	2	1
6	4	2
10	8	2
20	18	2
40	38	2
60	58	2
67	65	2
68	86	21
69	2	2
70	4	2
75	9	2

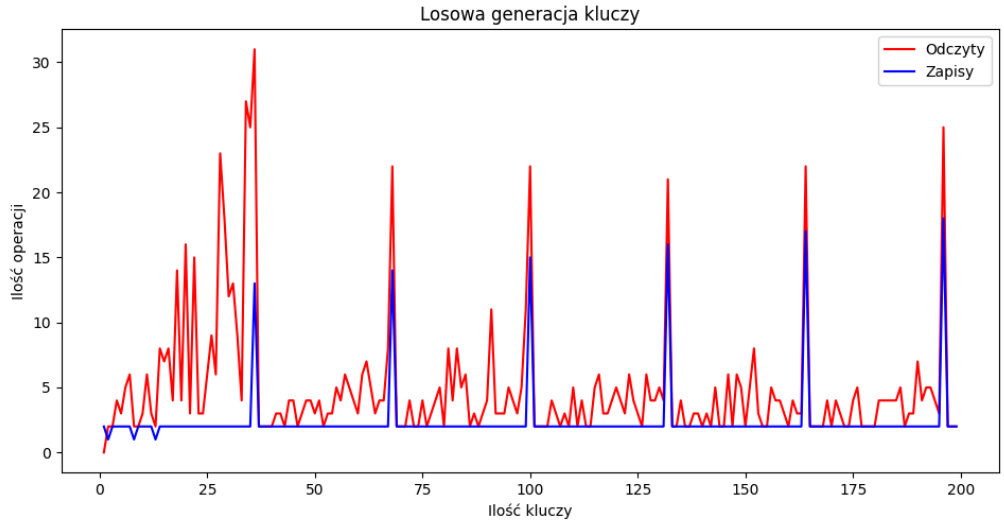
Tabela 1: Liczba odczytów i zapisów dla liniowo generowanych rekordów o limicie 5376



Rysunek 1: Liczba operacji odczytu i zapisu dla kolejno dodawanych rekordów o kluczach wygenerowanych liniowo o limicie 2688

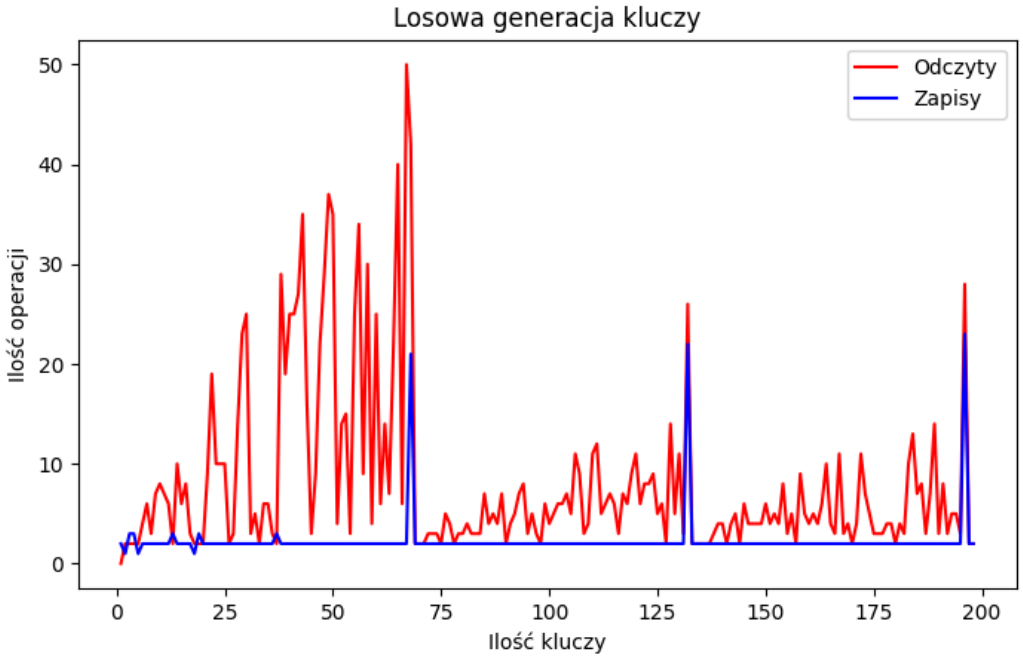


Rysunek 2: Liczba operacji odczytu i zapisu dla kolejno dodawanych rekordów o kluczach wygenerowanych losowo o limicie 2688

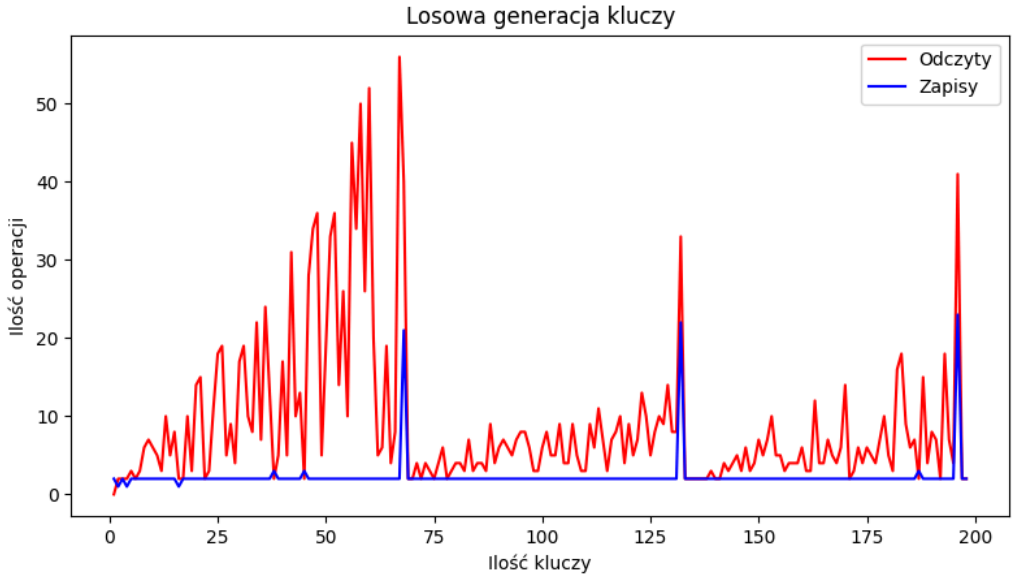


Rysunek 3: Liczba operacji odczytu i zapisu dla kolejno dodawanych rekordów o kluczach wygenerowanych losowo o limicie 2688

Rysunek 4: Liczba operacji odczytu i zapisu dla kolejno dodawanych rekordów o kluczach wygenerowanych liniowo o limicie 5376



Rysunek 5: Liczba operacji odczytu i zapisu dla kolejno dodawanych rekordów o kluczach wygenerowanych losowo o limicie 5376



Rysunek 6: Liczba operacji odczytu i zapisu dla kolejno dodawanych rekordów o kluczach wygenerowanych losowo o limicie 5376

Złożoność wstawiania, aktualizowania i znajdowania rekordu jest liniowo zależna od wielkości części nadmiarowej. Jest to spowodowane postacią części nadmiarowej - rekordy tam przechowywane są w strukturze przypominającej listę powiązaną, co powoduje, że przeszukiwanie i dodawanie elementów w odpowiednich miejscach nie jest wydajne. Widoczne są też nagłe wzrosty w liczbie odczytów i zapisów dla pewnych operacji, ponieważ bezpośrednio po wykonaniu operacji następuje reorganizacja pliku wymagająca dodatkowych operacji na dysku. Bezpośrednio po reorganizacji widoczny jest też spadek wymaganych operacji dyskowych. Jest to spowodowane faktem, że część nadmiarowa została połączona z częścią główną i plik indeksowy został zaktualizowany, co pozwala na szybsze znalezie-

nie pozycji danego rekordu. Liczba operacji zapisu jest praktycznie stała, poza operacjami, po których następuje reorganizacja pliku lub występuje przypadek graniczny, po którym trzeba zaktualizować klucz pierwszego rekordu pierwszej strony. Liczba operacji odczytu natomiast rośnie liniowo, z dużą wariancją dla kluczów generowanych losowo, ponieważ dane nie są dodawane rosnąco.

7. Podsumowanie

Cele projektu zostały zrealizowane. Aby zachować wydajność operacji na pliku, wymagana jest reorganizacja pliku, gdy część nadmiarowa osiąga rozmiar, który powoduje nieakceptowalne wzrost złożoności operacji.