Implementacja indeksowej organizacji pliku z użyciem B-drzewa

Patryk Miszke 193249

1. Opis implementacji metody:

Zasada buforowania w pamięci operacyjnej

Root drzewa w pamięci operacyjnej: Korzeń drzewa (root) jest trzymany w całości w pamięci RAM. Ma to na celu przyspieszenie dostępu do najważniejszego węzła, od którego rozpoczynają się wszystkie operacje na drzewie.

Pozostałe węzły w pliku binarnym: Każdy inny węzeł drzewa jest przechowywany w pliku binarnym. Węzły posiadają unikatowe indeksy, które informują, w którym miejscu w pliku się znajdują. Jest to możliwe przez zapisywanie ich dla tej samej wielkości: brakujące dane są zastępowane placeholderami do zapisu, natomiast przy odczycie danych do struktury w programie są pomijane.

Rekordy w pliku binarnym: Każdy rekord składa się z 6 doubli, które są kolejno parami 3 współrzędnych: x1, y1, x2, y2, x3, y3 oraz inta będącego kluczem (52 bajty na te 7 wartości). Rekordy w pliku głównym są przez drzewo umieszczane w prawidłowej kolejności, więc plik zawiera dane już posortowane względem ich kluczy.

Struktura węzła (Node)

Każdy węzeł drzewa zawiera następujące pola:

- **Klucze:** Lista kluczy przechowywanych w węźle.
- **Adresy kluczy:** Adresy w pliku głównym, które wskazują na dane skojarzone z kluczami.
- **Indeksy dzieci:** Indeksy węzłów dzieci w pliku binarnym.
- **Flaga `isLeaf`:** Wskazuje, czy węzeł jest liściem.
- **Unikatowy indeks:** Określa pozycję węzła w pliku binarnym (root ma wartość -1, ponieważ nie jest przechowywany w pliku).
- **Indeks rodzica:** Indeks węzła nadrzędnego, co umożliwia nawigację w górę drzewa.

Funkcjonalność

Program interaktywnie oferuje użytkownikowi różnorodne funkcje, które może wywoływać, aż do celowego zakończenia pracy programu (przycisk q):

- Dodawanie rekordów z pliku
- Dodawanie rekordu z klawiatury
- Odczytanie rekordu o danym kluczu
- Wyświetlenie struktury drzewa
- Wyświetlenie wszystkich rekordów z pliku głównego
- Usunięcie rekordów o kluczach wymienionych w pliku
- Usunięcie rekordu o kluczu wprowadzonym z klawiatury
- Zamiana rekordu o danym kluczu na nowy rekord

Po każdej operacji pokazywana jest liczba zapisów, odczytów oraz opcjonalnych operacji dyskowych wykorzystanych do realizacji funkcji. Wyświetlany odczyt oraz zapis dotyczy operacji niezbędnych do wykonania konkretnej funkcji. Operacje dodatkowe to te operacje, które są potrzebne do wprowadzonych usprawnień takich jak:

- **Index rodzica:** ułatwia poruszanie się po drzewie, jednak musimy aktualizować rodzica w każdym węźle, w którym się on zmieni

- **Wstawianie rekordów w odpowiedniej kolejności do pliku:** plik jest posortowany przez cały czas, co zmniejsza potrzebne odczyty do wyświetlenia wszystkich rekordów oraz umożliwia przesłanie osobie zewnętrznej już gotowego pliku bez całej struktury drzewa. Niestety wymaga to po każdym dodaniu i usunięciu rekordu aktualizacji wszystkich adresów rekordów w węzłach, które po nim występują.

Wykorzystanie zwalnianego miejsca

Struktura drzewa w programie zawiera wartość największego indeksu węzła, tak aby każdy kolejny miał następną unikatową wartość. Dodatkowo drzewo ma wektor z indeksami węzłów, które zostały usunięte. Dzięki temu w ich miejsce można wpisać te nowo utworzone, aby zaoszczędzić pamięć dyskową. Indeksy plików usuniętych mają priorytet nad tymi inkrementowanymi.

Funkcje zapisu, odczytu oraz usuwania rekordów

Zostały zaimplementowane zgodnie z algorytmami przedstawionym na wykładzie.

2. Specyfikacja formatu pliku testowego

   Pliki testowe muszą być w formacie txt dla ułatwienia wprowadzania do nich danych przez użytkownika.

   Plik testowy do zapisu: zawiera wypisanych po spacji 7 wartości, jest to 6 doubli (3 pary współrzędnych) oraz inta oznaczającego klucz. Każdy kolejny rekord jest zapisywany w osobnej linii.

   Plik testowy do odczytu: zawiera pojedynczego inta oznaczającego klucz rekordu, który program ma usunąć. Każdy kolejny klucz jest zapisany w osobnej linii.

3. Sposób prezentacji wyników działania programu

   B-drzewo:

```
Root: (62,14)
   |  |-- (25,4) (41,9)
   |  |  |  |-- (12,0) (14,1) (15,2) (24,3)
   |  |  |  |-- (31,5) (32,6) (33,7) (38,8)
   |  |     |-- (44,10) (52,11) (56,12) (61,13)
      |-- (83,19) (100,24) (127,29) (158,34)
      |  |  |-- (67,15) (73,16) (76,17) (78,18)
      |  |  |-- (85,20) (91,21) (98,22) (99,23)
      |  |  |-- (101,25) (123,26) (124,27) (125,28)
      |  |  |-- (128,30) (139,31) (141,32) (150,33)
      |     |-- (182,35) (190,36) (192,37) (200,38)
```

   Root jest wyświetlany jako pierwszy. Wartości to pary oddzielone spacją (klucz, adres klucza w pliku głównym). Każda kolejna gałęź to kolejna dziecko poczynając od lewej strony (najmniejszego dziecka) kończąc na dziecku z największymi wartościami.

   Rekord:

```
Record at position 0:
Values: 73.58 64.29 47.98 23.19 55.34 88.45 Key: 32
```

   Wszystkie rekordy:

```
Records from output file:
Values: 73.58 64.29 47.98 23.19 55.34 88.45  Key: 32
Values: 84.1 39.83 17.64 48.75 93.12 66.29  Key: 44
Values: 11.87 78.32 56.47 88.14 62.93 21.44  Key: 50
Values: 22.41 75.92 91.87 14.23 34.58 49.71  Key: 62
Values: 11.87 78.32 56.47 88.14 62.93 21.44  Key: 73
```

4. Eksperyment

Zależność stopnia drzewa:

Dodano dla danego stopnia drzewa 40 elementów, spisano ilość zapisów i odczytów do wykonania tej operacji oraz wyznaczono współczynnik alfa. Następnie usunięto 20 elementów i spisano liczbę zapisów oraz odczytów.

Stopień drzewa	Liczba rekordów	Zapis rekordów (40)	Usunięcie rekordów (20)	Współczynnik alfa
2	40	125 zapisów 296 odczytów	39 zapisów 877 odczytów	0,77

Stopień drzewa	Liczba rekordów	Zapis rekordów (40)	Usunięcie rekordów (20)	Współczynnik alfa
4	40	104 zapisów 218 odczytów	20 zapisów 347 odczytów	0,79

Stopień drzewa	Liczba rekordów	Zapis rekordów (40)	Usunięcie rekordów (20)	Współczynnik alfa
6	40	96 zapisów 179 odczytów	28 zapisów 360 odczytów	0,63

Wniosek:

Najbardziej optymalny stopień drzewa w przeprowadzonym eksperymencie to 4. Ma najlepszy współczynnik alfa oraz prawie najmniej operacji dyskowych potrzebnych do wykonania zapisów/usunięć. Można zaobserwować, że wraz ze wzrostem stopnia drzewa ilość zapisów oraz odczytów zazwyczaj maleje, ponieważ mamy mniej stron w pliku drzewa. Jednakże może wystąpić sytuacja, w której współczynnik alfa nie jest satysfakcjonujący, ponieważ stopień drzewa nie pozwala na optymalne rozmieszczenie rekordów w węzłach i marnujemy ich miejsce. Stopień 2 zaoferował lepszy współczynnik alfa, ale znacznie większe liczby operacji dyskowych w porównaniu do stopnia 6.

Zależność liczby rekordów dla B-drzewa stopnia 2:

Posiadając B-drzewo 2 stopnia wykonano operacje dodania 5 rekordów oraz usunięcia 5 rekordów. Spisano liczby odczytów oraz zapisów. Eksperyment wykonano na tym samym zestawie danych.

	Dodanie 5 rekordów		Usunięcie 5 rekordów	
Liczba rekordów	Zapis:	Odczyt:	Zapis:	Odczyt:
40	16	144	10	148
30	14	83	9	114
20	20	37	12	62
10	14	21	12	40

Wniosek:

Ilość zapisów dla operacji usuwania oraz dodawania jest bardzo podobna, ponieważ dodajemy te samą liczbę elementów, która nie jest duża, więc znacząco tego drzewa nie zmieniamy. Natomiast liczby odczytów znacznie się różnią i rosną wraz z ilością rekordów w B-drzewie. Jest to logiczne, gdyż im większe drzewo mamy - potrzebujemy większej ilości odczytów, aby się po nim poruszać, tym samym wykonywać operacje. Liczba operacji w przypadku 40 rekordów dla dodawania w porównaniu z 10 rekordami jest około 4,57 razy większa, natomiast dla usuwania około 3 razy większa.