

Katedra Inżynierii Oprogramowania
Wydział Elektroniki, Telekomunikacji i Informatyki
Politechnika Gdańska

STRUKTURY BAZ DANYCH

PROJEKT

Instrukcja

Opracowali:

prof. dr hab. inż. Krzysztof Goczyła
dr inż. Teresa Zawadzka
dr inż. Grzegorz Gołaszewski

Gdańsk, 2024

1. Cel projektu

Celem projektu jest wyrobienie praktycznej umiejętności w organizowaniu pamięci zewnętrznej bazy danych w struktury plikowe oraz w eksperymentowaniu na tych organizacjach.

2. Zadania do realizacji

W ramach projektu studenci wykonują dwa lub trzy zadania, których treść zamieszczono poniżej. Zadania należy zrealizować w języku ogólnego przeznaczenia (Pascal, C, C++, C# lub Java) i przesłać na platformę eNauczanie wersję źródłową programu. Po przesłaniu wymaganych plików należy przygotowane rozwiązanie przedstawić prowadzącemu. W ramach takiej prezentacji należy zaprezentować działanie programu zgodnie z podanymi wymaganiami oraz odpowiedzieć na pytanie(a) dotyczące kodu źródłowego. We wszystkich zadaniach programowi powinno towarzyszyć pisemne sprawozdanie, które, prócz nazwiska i imienia autora i treści zadania, powinno zawierać informacje specyficzne dla zadania (patrz: treści zadań). Sprawozdanie należy przesłać wraz z kodem źródłowym na platformę eNauczanie, jako osobny nie spakowany plik PDF.

Uwaga: We wszystkich zadaniach obowiązuje wersja algorytmu przedstawiona na wykładzie.

Zadanie 1.

Napisz program sortujący plik metodą scalania naturalnego lub scalania z użyciem wielkich buforów, lub metodą sortowania polifazowego. Taśmy powinny być zrealizowane w formie plików dyskowych. Należy zaimplementować symulację odczytu oraz zapisu blokowego jako oddzielną warstwę logiki programu. Warstwa ta powinna udostępniać warstwie sortowania co najmniej 2 operacje: odczytu oraz zapisu pojedynczego rekordu. Rekordy do posortowania powinny być generowane zarówno losowo, jak i z klawiatury. Program powinien wyświetlić zawartość pliku przed posortowaniem i po posortowaniu. Program powinien też dawać możliwość wyświetlenia pliku po każdej fazie sortowania. Na zakończenie należy wyświetlić liczbę faz sortowania oraz liczbę odczytów i zapisów stron na dysk. Dodatkowo program powinien dawać możliwość wczytywania danych testowych z pliku testowego.

Na tak przygotowanym programie należy przeprowadzić eksperyment, w którym:

- określona zostanie liczba faz sortowania oraz liczba odczytów i zapisów stron dyskowych w zależności od liczby losowanych rekordów (dla min. 5 znacznie różniących się od siebie liczb sortowanych rekordów),
- wyznaczona zostanie teoretyczna liczba faz oraz operacji dyskowych dla badanych liczb sortowanych rekordów,
- uzyskane wyniki przedstawione zostaną na wykresie funkcji (należy zwrócić uwagę na zachowanie skali na obu osiach wykresu {może to być skala liniowa lub logarytmiczna}),
- porównane zostaną uzyskane wyniki praktyczne i teoretyczne, tu należy również spróbować wyjaśnić dostrzeżone rozbieżności.

Uwaga: Typ rekordu pliku należy pobrać u prowadzącego projekt.

Sprawozdanie ma zawierać:

- krótki opis zastosowanej metody (w szczególności – liczbę zastosowanych taśm),
- specyfikację formatu pliku testowego,
- opis sposobu prezentacji wyników działania programu,
- opis eksperymentu.

Termin realizacji zadania: 7 tygodni od rozpoczęcia semestru.

Maksymalna punktacja za poszczególne elementy zadania wynosi:

- Poprawna implementacja wybranego algorytmu:
 - **20 punktów** przy implementacji scalania naturalnego,

- **25 punktów** przy implementacji scalania z użyciem wielkich buforów lub sortowania polifazowego.
- Przygotowanie sprawozdania – **2 punkty**.
- Przeprowadzenie eksperymentu dotyczącego zaimplementowanej metody oraz opisanie go w sprawozdaniu – **8 punktów**,
w opisie tym powinny znaleźć się:
 - opis konstrukcji/sposobu przeprowadzenia eksperymentów,
 - wyniki liczbowe eksperymentów wykres łącznej liczby odczytów rekordów/stron dyskowych i łącznej liczby zapisów rekordów/stron dyskowych w funkcji N ,
 - komentarz do wyników uwzględniający w szczególności porównanie wyników praktycznych z teoretycznymi.

Zadanie 2.

Zaprojektuj i zaimplementuj wybraną indeksową organizację pliku (B-drzewo, B+-drzewo lub indeksowo-sekwencyjną). Zrealizuj operacje: wstawiania rekordu, odczytu rekordu, przeglądania całej zawartości pliku i indeksu zgodnie z kolejnością wartości klucza oraz operację reorganizacji pliku (jeśli wybrana organizacja tego wymaga). Operacja reorganizacji powinna być wywoływana automatycznie w miarę potrzeb oraz powinna istnieć możliwość reorganizacji na żądanie. Typ rekordu pliku przyjmij jako typ rekordu z zadania 1 rozszerzony o klucz (liczbę naturalną) [oznacza to, że podane w rekordzie kryterium sortowania przestaje obowiązywać, a rekordy są sortowane zgodnie z wartością klucza]. Po każdej przeprowadzonej operacji podawaj liczbę faktycznie zrealizowanych operacji odczytu lub zapisu stron dyskowych. Program powinien dawać możliwość wyświetlania zawartości pliku z danymi i indeksu po każdej operacji zmieniającej zawartość pliku (tj. po wstawieniu, aktualizacji lub usunięciu rekordu) [wyświetlanie zawartości pliku oznacza przedstawienie w sposób czytelny dla człowieka wewnętrznej struktury tych plików. Dla B-drzewa wyświetlona zawartość indeksu powinna dać możliwość ustalenia budowy B-drzewa oraz rozmieszczenia w nim kluczy, a dla struktury indeksowo-sekwencyjnej wyświetlenie zawartości pliku z danymi powinno dać możliwość sprawdzenia, czy rekord znajduje się w części głównej, czy nadmiarowej, dokładnie w którym miejscu, ile jest pustych miejsc na poszczególnych stronach pliku, czy też zawartości poszczególnych łańcuchów przepełnień.] Dodatkowo program powinien dawać możliwość wczytywania danych testowych z pliku testowego. Danymi testowymi powinna być dowolna sekwencja operacji wstawiania, aktualizacji i usuwania rekordu (jeśli zostały zaimplementowane). Program nie czytający poleceń z pliku testowego powinien działać w sposób interaktywny, tzn. po pobraniu komendy wykonać ją, przedstawić wynik jej wykonania, po czym czekać na następną komendę.

Na tak stworzonym programie przeprowadź eksperymenty mające na celu ustalenie wpływu parametrów implementacyjnych (stopień drzewa, współczynnik alfa, warunek wykonania automatycznej reorganizacji) na złożoność poszczególnych operacji w pliku (wstawianie, usuwanie...). Poza parametrami implementacyjnymi w eksperymencie uwzględnij również liczbę przechowywanych w pliku rekordów. Omów uzyskane wyniki. Czy istnieje optymalna wartość badanych parametrów?

Uwaga: na potrzeby prezentacji należy stopień drzewa ustawić na wartość $d=2$ (dla B-drzewa oraz B+-drzewa) lub współczynnik blokowania pliku z danymi na $b=4$ (dla struktury indeksowo-sekwencyjnej).

Sprawozdanie powinno zawierać:

- informację, jaką metodę indeksowania przyjęto do implementacji,
- opis implementacji przyjętej metody (zasady buforowania w pamięci operacyjnej, ograniczenia, ew. usprawnienia itp.),
- specyfikacja formatu pliku testowego,
- opis sposobu prezentacji wyników działania programu,
- opis przeprowadzonych eksperymentów.

Termin realizacji zadania: 11 tygodni od rozpoczęcia semestru.

Punktacja:

- Poprawna implementacja wybranego algorytmu - **20 punktów**.
- Przygotowanie sprawozdania – **2 punkty**
- Przeprowadzenie eksperymentu dotyczącego zaimplementowanej metody oraz opisanie go w sprawozdaniu – **8 punktów**,
w opisie tym powinny znaleźć się:
 - opis eksperymentów (konstrukcja oraz sposób przeprowadzenia),
 - wyniki przeprowadzonych eksperymentów (średnia liczba odczytów i zapisów stron dyskowych na jedną operację, zajętość pamięci przez plik i indeks w funkcji liczby rekordów i in.),
 - komentarz na temat implementacji i uzyskanych wyników.
- Zaimplementowanie w wybranym algorytmie operacji usuwania i aktualizacji rekordu – **10 punktów**.
- Zaimplementowanie mechanizmu pozwalającego na ponowne wykorzystanie zwalnianego przy usuwaniu miejsca (tylko dla B-drzewa i B+drzewa). – **5 punktów**.

Zadanie 3. – dodatkowe

Zaprojektuj i zaimplementuj organizację pliku opartą na rozpraszaniu dynamicznym (rozszerzalnym lub liniowym). Zrealizuj operacje: wstawiania rekordu, aktualizacji rekordu, odczytu rekordu, usuwania rekordu oraz przeglądania całej zawartości pliku. Typ rekordu pliku przyjmij ten sam co w zadaniu 2. Po każdej przeprowadzonej operacji podawaj liczbę faktycznie zrealizowanych operacji odczytu lub zapisu stron dyskowych. Program powinien dawać możliwość wyświetlania zawartości pliku i skorowidza (dla metody rozszerzalnej) po każdej operacji zmieniającej zawartość pliku (tj. po wstawieniu, aktualizacji lub usunięciu rekordu). Dodatkowo program powinien dawać możliwość wczytywania danych testowych z pliku testowego. Danymi testowymi powinna być dowolna sekwencja operacji wstawiania, aktualizacji i usuwania rekordu.

Sprawozdanie powinno zawierać:

- informację, jaką metodę rozpraszania przyjęto do implementacji,
- opis implementacji przyjętej metody (zasady buforowania w pamięci operacyjnej, ograniczenia, usprawnienia itp.),
- specyfikację formatu pliku testowego,
- opis sposobu prezentacji wyników działania programu.

Termin oddania opracowania: 13 tygodni od rozpoczęcia semestru.

Punktacja: **30 punktów** za wykonanie zadania.

3. Zasady oceniania

Warunkiem zaliczenia laboratorium jest zaliczenie pierwszych dwóch zadań. Minimum potrzebne do zaliczenia zadań pierwszego i drugiego to odpowiednio 16 i 22 punkty. W zadaniu drugim nie jest wymagane zaimplementowanie operacji usuwania i aktualizacji rekordu, a także w przypadku B-drzewa oraz B+drzewa mechanizmu ponownego wykorzystania zwalnianego miejsca. Skutkować to będzie jednak niższą punktacją.

Oddawanie zadań ze spóźnieniem jest możliwe dla pierwszych dwóch zadań, jednak **opóźnienie o każdy rozpoczęty tydzień powoduje odjęcie 5 punktów**. Ponadto spóźnienie może wynieść **maksymalnie 2 tygodnie**. Zadania 3 nie można oddawać ze spóźnieniem.

Dodatkowo **zadanie należy zaprezentować** prowadzącemu **w przeciągu dwóch tygodni** od daty przesłania zadania na eNauczanie.

Przy ocenianiu zadań brana jest pod uwagę jakość przyjętego rozwiązania, liczba i sposób przeprowadzenia eksperymentów oraz treść sprawozdania. Sposób realizacji interfejsu użytkownika ma mniejsze znaczenie.

Przy wyznaczaniu oceny 75 punktów stanowi 100%. Maksymalna liczba punktów do zdobycia wynosi 110.

W razie niezaliczenia któregoś z obowiązkowych zadań możliwe jest jego poprawianie po uprzednim uzgodnieniu z prowadzącym warunków owej poprawy. Poprawiając można zdobyć minimalną liczbę punktów za zadanie (odpowiednio 16 i 22). Poprawiać zadania można do końca semestru (należy się rozliczyć przed rozpoczęciem sesji podstawowej).