Wydział Informatyki i Telekomunikacji
Politechnika Wrocławska

Analiza eksperymentalna nowych algorytmów sortowania w miejscu

Damian Baliński
Nr indeksu: 250332

Praca inżynierska napisana
pod kierunkiem
dr inż. Zbigniewa Gołębiewskiego

Politechnika
Wrocławska
Wrocław 2021

Spis treści

Wstęp

Ogólna historia algorytmów sortujących. Motywacja tworzenie algorytmów

1.1 Cel pracy

TODO: Motywem przewodnim pracy jest analiza nowoczesnych algorytmów sortujących w miejscu, takich jak koncepcja QuickMerge Sort. W tym celu przygotowano analizę porównawczą podstawowych algorytmów sortujących oraz dokonano przeglądu zmodyfikowanych wersji tych algorytmów oraz przeanalizowano nowoczesne algorytmy hybrydowe, będące połączeniem dwóch lub wielu algorytmów podstawowych.

1.2 Zakres pracy

TODO: Aby ułatwić analizę algorytmów przygotowany został silnik testujący oraz silnik graficzny, które w oparciu o plik konfiguracyjny przeprowadzają testy oraz tworzą wizualizację wyników tych testów. Wykorzystując podane narzędzia została przeprowadzona analiza podstawowych oraz hybrydowych algorytmów.

1.3 Przegląd literatury

TODO: Ogólny opis pracy Sebastiana Wilda. Pomysł na algorytm QuickMerge Sort.

1.4 Zawartości pracy

TODO: Ogólny sposób organizacji dokumentu. Rozdział pierwszy - analiza matematyczna problemu. Rozdział drugi - przegląd podstawowych algorytmów sortujących. Rozdział trzeci - przegląd hybrydowych algorytmów sortujących.

Analiza problemu

2.1 Model matematyczny przypadku średniego

2.2 Model matematyczny przypadku pesymistycznego

2.3 Założenia

2.3.1 Założenia odnośnie testowanych parametrów

TODO: Testowana będzie złożoność czasowa, w tym celu analizowane są takie parametry jak: liczba porównań, liczba swapów oraz liczba przypisań, z pominięciem operacji wykonywanych na iteratorach pętli - wyjaśnienie dlaczego.

TODO: Ponieważ rzeczywisty czas wykonywania algorytmu różni się w zależności od maszyny oraz architektury systemu na którym przeprowadzany test, zostało przyjęte następujące założenie: Złożoność czasowa została określona wzorem:

$$\boldsymbol{T = n_c + 3 \cdot n_s + n_a}$$

Gdzie:
$\boldsymbol{T}$ - czas trwania algorytmu
$\boldsymbol{n_c}$ - liczba operacji porównania
$\boldsymbol{n_s}$ - liczba operacji zamiany miejsc
$\boldsymbol{n_a}$ - liczba operacji przypisania

2.3.2 Założenia odnośnie danych wejściowych

TODO: Wiele algorytmów sortujących bazuje na pewnych założeniach odnośnie wejściowego zbioru danych. Np. algorytm ... doskonale radzi sobie ze zbiorem danych prawie posortowanym, tzn. takim w którym ... . W tej pracy zakładamy że dane wejściowe będą losowym ciągiem liczb powtórzeniami.

Przegląd podstawowych algorytmów sortujących

3.1 Quick Sort

algorytm stabilny	**NIE**
algorytm miejscowy	**TAK**
optymistyczna złożoność czasowa	$\boldsymbol{O(n \log n)}$
pesymistyczna złożoność czasowa	$\boldsymbol{O(n^2)}$
średnia złożoność czasowa	$\boldsymbol{O(n \log n)}$
złożoność pamięciowa	$\boldsymbol{O(1)}$

Historia algorytmu Quick Sort sięga drugiej połowy XX wieku. W roku 1959 brytyjski naukowiec Tony Hoare opracował, a dwa lata później opublikował pierwszą wersję tego algorytmu. Od tamtego czasu powstało wiele udoskonaleń tego algorytmu, jednak jego koncepcja nadal jest widoczna we współczesnych językach programowania [1]. Na cześć algorytmu Quick Sort standardowa funkcja sortująca w języku C++ nosi nazwę qsort [2].

Algorytm Quick Sort składa się z dwóch etapów. Pierwszym z nich jest partycjonowanie zbioru wejściowego. Po tym kroku tablica wejściowa jest rozbita na dwa rozłączne zbiory, w których wszystkie elementy pierwszego zbioru są skumulowane po lewej stronie tablicy oraz każdy z tych elementów jest większy od dowolnego elementu z drugiej tablicy. Drugim etapem jest rekurencyjne sortowanie lewej oraz prawej podtablicy. Algorytm Quick Sort wykorzystuje technikę dziel i zwyciężaj, ponieważ problem sortowania tablicy wejściowej rozbija na sortowanie dwóch podtablic.

3.1.1 Analiza algorytmu Quick Sort

Liczba operacji wykonywanych przez algorytm Quick Sort została przeanalizowana pod kątem trzech przypadków: optymistycznego, średniego oraz pesymistycznego.

Dla algorytmu Quick Sort przypadek optymistyczny (3.1) następuje wówczas, gdy algorytm partycjonowania przy każdym wywołaniu dzieli tablicę wejściową na dwie równe części. Efekt ten uzyskano, wprowadzając dane już posortowane oraz stosując algorytm wybierający pivot dokładnie w połowie tablicy. Z analizy eksperymentalnej wynika, że w przypadku optymistycznym algorytm działa ze złożonością czasową $\boldsymbol{O(n \log n)}$.

[1] Dokumentacja biblioteki sortującej w języku java: https://docs.oracle.com/en/java/javase/11/docs/api/java.base/java/util/Arrays.html

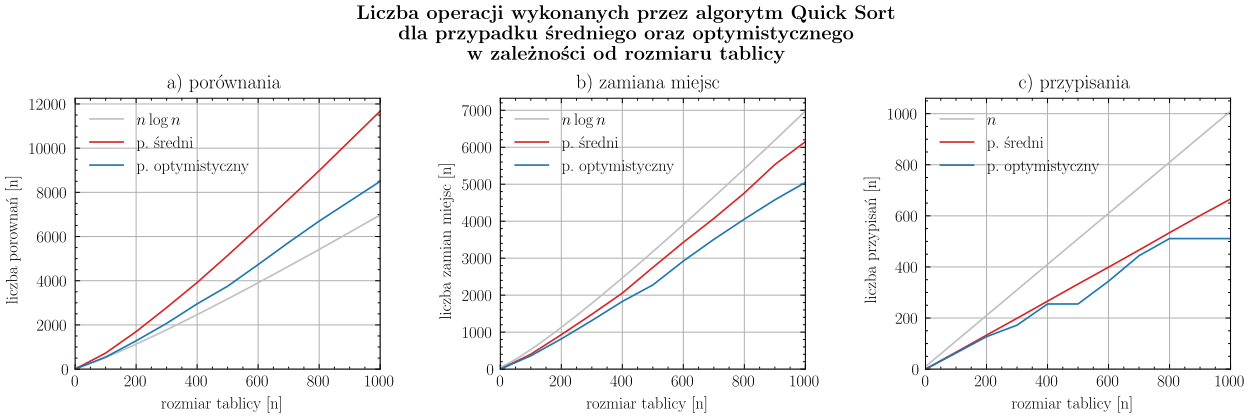
[2] Dokumentacja funkcji sortującej qsort: https://en.cppreference.com/w/cpp/algorithm/qsort

Przypadek pesymistyczny (3.2) zachodzi, gdy drzewo wowołań rekurencyjnych jest możliwie najgłębsze. Efekt ten uzyskano, wprowadzając dane już posortowane oraz stosując algorytm wybierający pivot jako ostatni element tablicy. W tej sytuacji w kolejnych iteracjach rozpatrywana jest tablica z rozmiarem o jeden mniejszy od poprzedniej, a więc drzewo wywołań rekurencyjnych ma głębokość $\boldsymbol{n}$. Z analizy wynika, że złożoność czasowa algorytmu w przypadku pesymistycznym wynosi $\boldsymbol{O(n^2)}$.

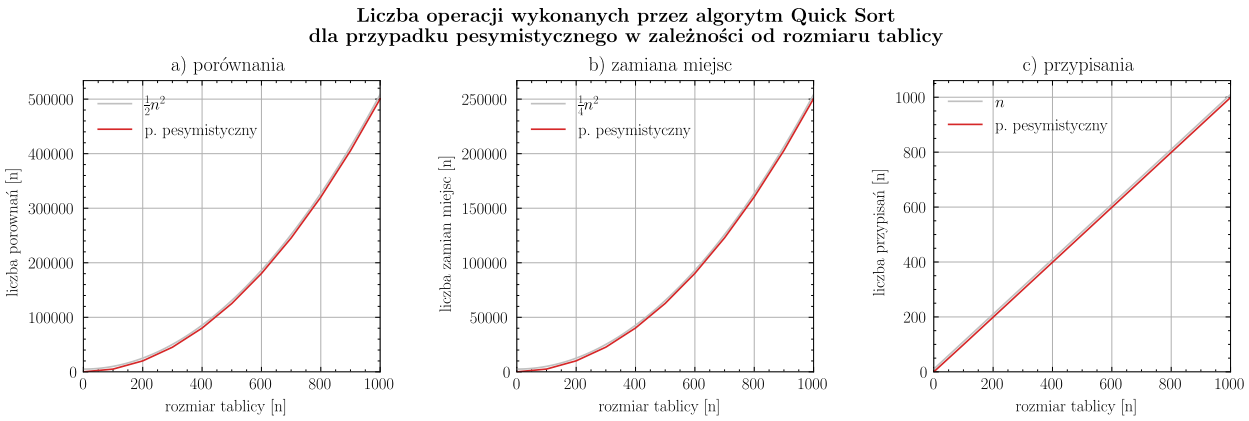
Przypadek średni (3.1) został zbadany wprowadzając losowe dane z powtórzeniami oraz stosując algorytm wybierający pivot jako ostatni element tablicy. Analiza eksperymentalna wykazała, że w przypadku średnim algorytm Quick Sort ma złożoność czasową równą $\boldsymbol{O(n \log n)}$, a więc jest tego samego rzędu co dla przypadku optymistycznego.

Porównując liczbę wykonywanych operacji można zauważyć, że algorytm Quick Sort wykonuje prawie dwa więcej operacji porównania niż operacji zamiany miejsc. Liczba pojedynczych operacji przypisania rośnie liniowo, a więc jest znikoma w porównaniu z liczbą pozostałych operacji.

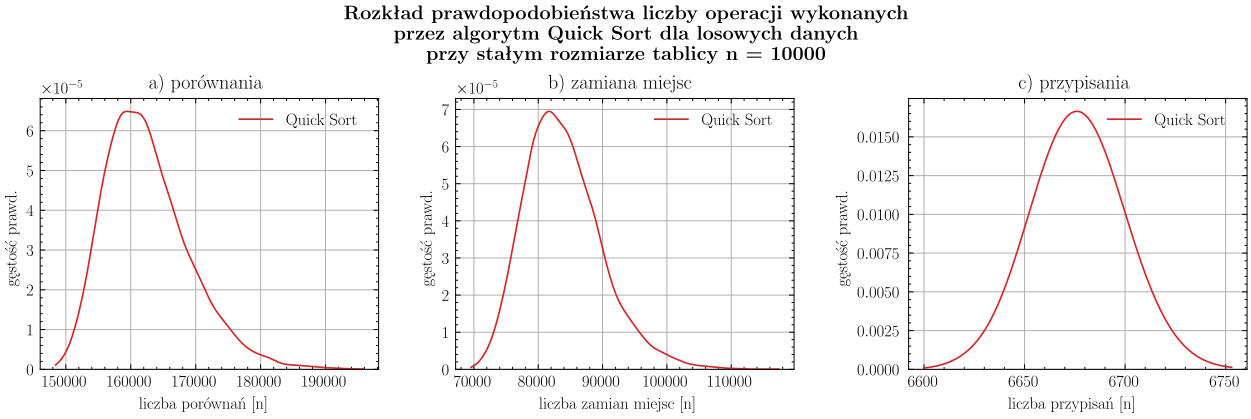
Analizując rozkład prawdopodobieństwa liczby wykonanych operacji (3.3) użyto tablicy losowych danych o stałym rozmiarze $\boldsymbol{n = 10000}$. Można zauważyć, że liczba operacji porównania oraz liczba operacji zamiany miejsc nie są przedstawiane za pomocą rozkładu normalnego. Bardziej prawdopodobne jest wykonanie większej liczby tych operacji w stosunku do wartości średniej. Z kolei rozkład liczby wykonanych operacji przypisania przedstawia się za pomocą rozkładu normalnego, z jednakowym prawdopodobieństwem liczba ta może być większa lub mniejsza od wartości średniej.



Rysunek 3.1



Rysunek 3.2



Rysunek 3.3

3.1.2 Problemy związane z algorytmem Quick Sort

Głównym problemem algorytmu Quick Sort jest jego słaba pesymistyczna złożoność czasowa. Ponieważ algorytm działa rekurencyjnie, w przypadku pesymistycznym głębokość drzewa wywołań rekurencyjnych może przekroczyć maksymalną liczbę ramek stosu, powodując awaryjne zatrzymanie programu.

Kolejnym problemem tego algorytmu jest stosunkowo duża liczba wykonywanych operacji porównania w stosunku do liczby pozostałych operacji. Punkt ten jest szczególnie istotny w sytuacji, gdy sortowane są złożone struktury, dla których wykonanie pojedynczej operacji porównania jest znacznie kosztowniejsze od pozostałych operacji. W tym przypadku bardziej wskazanym wydaje się użycie algorytmu Merge Sort, którego analizę przeprowadzono w kolejnym rozdziale.

3.1.3 Możliwości optymalizacyjne

Ponieważ złożoność czasowa algorytmu Quick Sort uwarunkowana jest poprzez złożoność algorytmu partycjonowania, optymalizacja algorytmu może opierać się ulepszanie algorytmu partycjonowania, aby jak najefektywniej wyszukiwał pivot, zapewniając podział tablicy wejściowej na dwie tablice o zbliżonej długości.

3.2 Merge Sort

algorytm stabilny	**TAK**
algorytm miejscowy	**NIE**
optymistyczna złożoność czasowa	$\boldsymbol{O(n \log n)}$
pesymistyczna złożoność czasowa	$\boldsymbol{O(n \log n)}$
średnia złożoność czasowa	$\boldsymbol{O(n \log n)}$
złożoność pamięciowa	$\boldsymbol{O(n)}$

Algorytm Merge Sort został opracowany przez Johna von Neumanna w 1945 roku. Tak jak Quick Sort, algorytm ten wykorzystuje technikę dziel i zwyciężaj aby rekurencyjnie rozbić problem sortowania danych wejściowych na dwie listy mniejszej długości. W przeciwieństwie do algorytmu Quick Sort, algorytm Merge Sort do poprawnego działania potrzebuje dodatkowej pamięci. W rozważanej implementacji algorytm alokuje dodatkowy bufor długości $\boldsymbol{n/2}$.

Algorytm Merge Sort składa się z trzech etapów. Pierwszym krokiem jest podział tablicy wejściowej na dwie części o zbliżonej długości. Drugim etapem jest rekurencyjne sortowanie każdej z części. Ostatnim krokiem jest scalenie tablic cząstkowych zgodnie z porządkiem sortowania. Aby efektywnie wykonać ostatni krok, algorytm Merge Sort potrzebuje zewnętrznego bufora.

3.2.1 Analiza algorytmu Merge Sort

Liczba operacji wykonywanych przez algorytm Merge Sort została przeanalizowana pod kątem trzech przypadków: optymistycznego, średniego oraz pesymistycznego.

Dla algorytmu Merge Sort przypadek optymistyczny (3.4) następuje wówczas, gdy w każdym kroku scalania lewa podtablica zawiera tylko elementy mniejsze od wszystkich elementów z prawej podtablicy. Efekt ten uzyskano, wprowadzając dane już posortowane. Z analizy eksperymentalnej wynika, że w tym przypadku algorytm Merge Sort działa ze złożonością czasową rzędu $\boldsymbol{O(n \log n)}$.

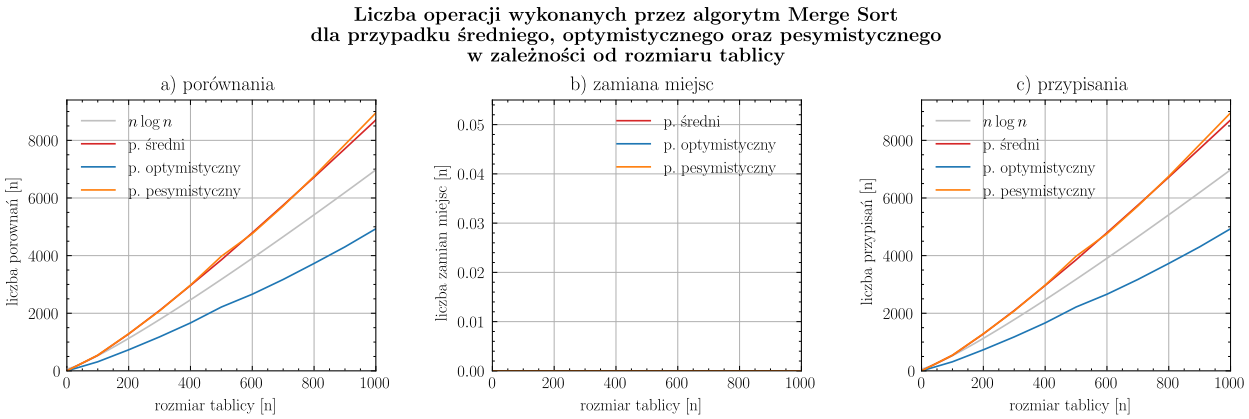
Przypadek pesymistyczny (3.4) zachodzi, gdy liczba porównać między sobą elementów w trakcie scalania jest możliwie największa. Efekt ten uzyskano, preparując dane wejściowe w taki sposób, aby w każdym kroku scalania porównywane tablice zawierały elementy na przemian większe oraz mniejsze. Z analizy wynika, że złożoność czasowa algorytmu w przypadku pesymistycznym jest równa $\boldsymbol{O(n \log n)}$, a więc jest tego samego rzędu co złożoność optymistyczna.

Przypadek średni (3.4) zbadano wprowadzając na wejście losowe dane z powtórzeniami. Analiza wykazała, że w przypadku średnim złożoność algorytmu jest równa $\boldsymbol{O(n \log n)}$.

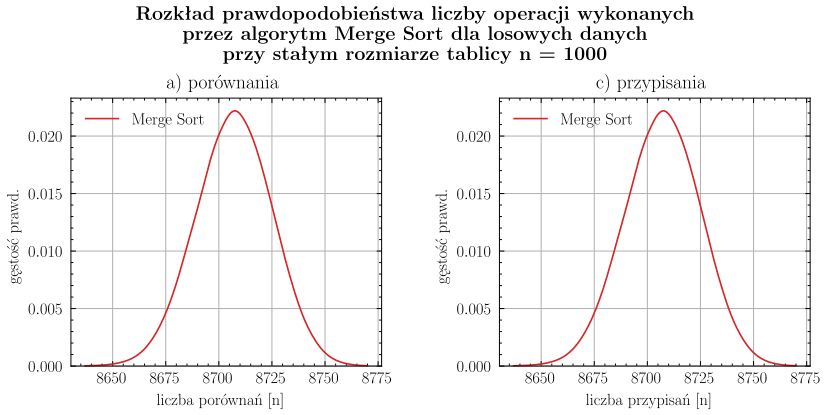
W przeciwieństwie do algorytmu Quick Sort, algorytm Merge Sort wykonuje dokładnie tyle samo operacji porównania co operacji przypisania. Jest to zgodne z rzeczywistością, ponieważ w trakcie scalania każdej operacji porównania towarzyszy przypisanie wartości do bufora. Można zauważyć, że algorytm Merge Sort w rozpatrywanej postaci nie wykonuje operacji zamiany miejsc.

Do analizy rozkładu prawdopodobieństwa liczby wykonanych operacji (3.5) użyto tablicy losowych danych o stałym rozmiarze $\boldsymbol{n = 1000}$. W analizie pominięto liczbę operacji przypisania, która jest zerowa dla każdego przypadku. Analiza eksperymentalna dowodzi, że w przeciwieństwie do algorytmu Quick Sort, liczba operacji

porównania dla algorytmu Merge Sort tworzy rozkład normalny, a więc prawdopodobieństwa otrzymania większej oraz mniejszej liczby operacji przypisania są jednakowe. Ponieważ dla badanego algorytmu, liczba operacji przypisania jest równa liczbie operacji porównania, rozkłady tych wartości są jednakowe.



Rysunek 3.4



Rysunek 3.5

3.2.2 Problemy związane z algorytmem Merge Sort

Głównym problemem algorytmu Merge Sort jest konieczność alokacji dodatkowego bufora pamięci. W rozważanej implementacji algorytm alokuje dodatkowy bufor długości $\boldsymbol{n/2}$, a więc może okazać się bezużyteczny dla systemów z ograniczonym zasobem pamięci.

3.2.3 Możliwości optymalizacyjne

Jednym ze sposobów na optymalizację algorytmu Merge Sort jest próba przekształcenia go w algorytm działający w miejscu. Cel ten może zostać osiągnięty poprzez skrzyżowanie algorytmu Merge Sort z innym algorytmem sortującym w taki sposób, aby bufor dodatkowej pamięci był częścią tablicy wejściowej. Rozwiązanie to zostanie przeanalizowane w dalszych rozdziałach.

Przegląd hybrydowych algorytmów sortujących

4.1 Główne sposoby modyfikacji algorytmów

1. Podmiana algorytmów składowych, np. inny algorytm partycjonowania.
2. Łączenie wielu algorytmów w jeden.

4.2 Rodzina algorytmów Quick Sort z deterministycznym algorytmem wyboru pivota

Algorytmy Quick Sort z różnymi algorytmami partycjonowania oraz różnymi algorytmami wyboru pivota Wykresy porównujące algorytmy.
1. Partycjonowanie metodą Lemuto (domyślne)
2. Partycjonowanie metodą Hoare
3. Wybór pivota metodą median of three - pierwszy, środkowy, ostatni.
4. Wybór pivota jako mediana of medians
5. Wybór pivota metodą pseudomedian of nine
6. Wybór pivota algorytmem quick select

4.3 Rodzina algorytmów Quick Sort z niedeterministycznym algorytmem wyboru pivota

1. QuickSort z losowaniem pivota
2. QuickSort z losowaniem trzech liczb, wybór mediany (power of three choices)

4.4 QuickMerge Sort

Koncepcja algorytmu, mocne strony (Merge Sort bez konieczności alokacji pamięci)

4.4.1 Pseudokod

4.4.2 Analiza algorytmu

Wykresy liczby wykonywanych operacji w porównaniu do algorytmów bazowych. Wykresy gęstości liczby wykonywanych operacji dla stałej liczby n, np n = 10000.

4.4.3 Wnioski

Wyniki analizy porównawczej

4.5 Intro Sort

Ogólny opis algorytmu, gdzie jest wykorzystywany (std::sort w g++), zalety.

4.5.1 Pseudokod

4.5.2 Analiza algorytmu

Wykresy liczby wykonywanych operacji w porównaniu do algorytmów bazowych. Wykresy gęstości liczby wykonywanych operacji dla stałej liczby n, np n = 10000. Wykresy dla różnych algorytmów partycjonowania.

4.5.3 Wnioski

Wyniki analizy porównawczej

Implementacja systemu

5.1 Struktura systemu

Aplikacja składa się z dwóch modułów: silnika testującego oraz silnika graficznego. Działanie systemu jest określane na podstawie współdzielonego pliku konfiguracyjnego. W pliku konfiguracyjnym określone są rodzaje testów jakie należy przeprowadzić oraz metadane potrzebne do wygenerowania wizualizacji.

Silnik testujący to generyczna biblioteka algorytmów sortujących oraz narzędzie przetwarzające te algorytmy. Aplikacja w oparciu o plik konfiguracyjny generuje zestaw testowy oraz utrwala wyniki przeprowadzonych testów na dysku. W zależności od konfiguracji, silnik testujący może sumować, zliczać lub uśredniać liczbę wykonywanych operacji takich jak: liczba porównań, liczba operacji zamiany miejsc, liczba operacji przypisania oraz czas trwania algorytmu. Ta część aplikacji została napisana w języku C++ [1] z wykorzystaniem technik programowania obiektowego.

Silnik graficzny to zbiór skryptów przetwarzających wyniki z silnika testującego. Na podstawie pliku konfiguracyjnego oraz danych testowych generowane są wizualizacje graficzne w postaci wykresów, dzięki czemu użytkownik końcowy może w łatwy sposób analizować oraz porównywać badane algorytmy. Ta część systemu została napisana w języku Python [2] przy użyciu biblioteki matplotlib [3].

5.2 Koncepcje architektury silnika testującego

5.2.1 Wstrzykiwanie zależności

Większość algorytmów sortujących składa się z kilku odrębnych kroków. Niektóre z tych kroków są na tyle złożone, że stanowią osobne algorytmy. Dla przykładu jednym etapów sortowania metodą Quick Sort jest partycjonowanie danych wejściowych na rozłączne zbiory. Aby w łatwy sposób umożliwić modyfikację testowanych algorytmów, bez konieczności ponownej implementacji całego procesu, zastosowano technikę wstrzykiwania zależności. Jeżeli algorytm testujący korzysta z innego algorytmu, to algorytm składowy jest wstrzykiwany w trakcie działania programu. Dzięki temu lekka modyfikacja testowanego algorytmu ogranicza się do podmiany jego algorytmów składowych, bez konieczności ingerowania w strukturę bazową.

5.2.2 Obiektowość

Aby uprościć organizację kodu zastosowano model obiektowy. Każdy z algorytmów wykorzystywanych w systemie został zamodelowany za pomocą odrębnej klasy. Dla każdej rodziny algorytmów tego samego typu istnieje nadrzędna klasa bazowa określająca interfejs dla tej rodziny. Korzyści wynikające z zastosowanego modelu, takie jak statyczny polimorfizm oraz dziedziczenie, gwarantują bardziej wiarygodne działanie programu oraz umożliwiają wykrywanie błędów strukturalnych już na etapie kompilacji projektu.

[1] Dokumentacja języka C++: https://en.cppreference.com
[2] Dokumentacja języka Python: https://docs.python.org/3/
[3] Dokumentacja biblioteki matplotlib: https://matplotlib.org/

5.2.3 Bezstanowość

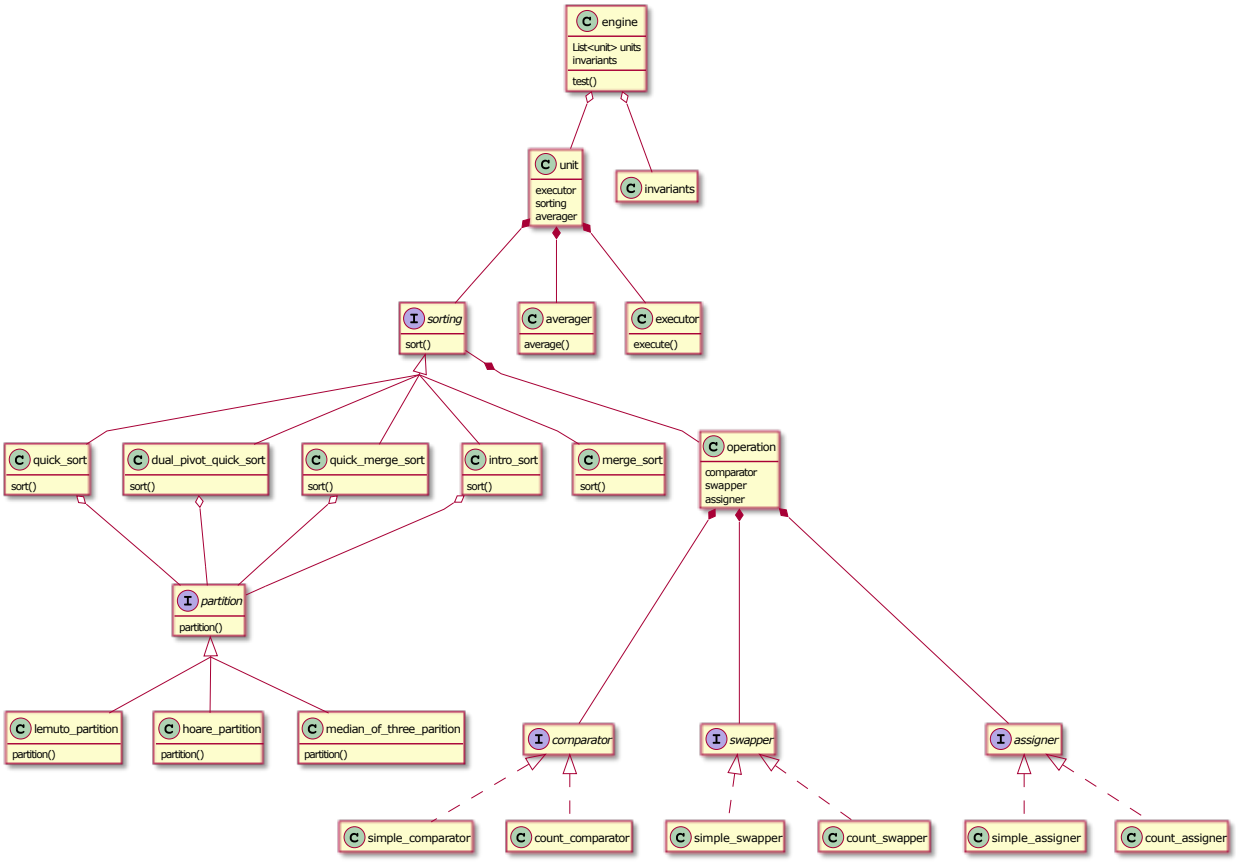
Powszechnym problemem w programowaniu obiektowym jest przechowywanie stanu. Problem ten wynika po części z praktyki hermetyzacji danych wewnątrz obiektowej abstrakcji. Użytkownik zewnętrzny korzystając z interfejsu danego komponentu nie ma dostępu do procesów zachodzących w jego wnętrzu. Może to prowadzić do tzw. efektów ubocznych (ang. side effects), przez co wyniki zwracane przez program stają się niewiarygodne.

Aby tego uniknąć zastosowano model bezstanowy. Żaden z algorytmów sortujących w zaimplementowanym systemie nie posiada zmiennych składowych, które mogły by zostać zmodyfikowane w trakcie działania programu. Podczas testowania dane są przekazywane poprzez sygnatury metod, wzorując się na technice programowania funkcyjnego. Dzięki temu wszystkie algorytmy sortujące wykorzystane w implementacji są oznaczone jako niemodyfikowalne, nie mogą zmienić stanu aktualnie testowanego algorytmu. Gwarantuje to całkowitą separację poszczególnych testów.

5.3 Model aplikacji

5.3.1 Diagram klas

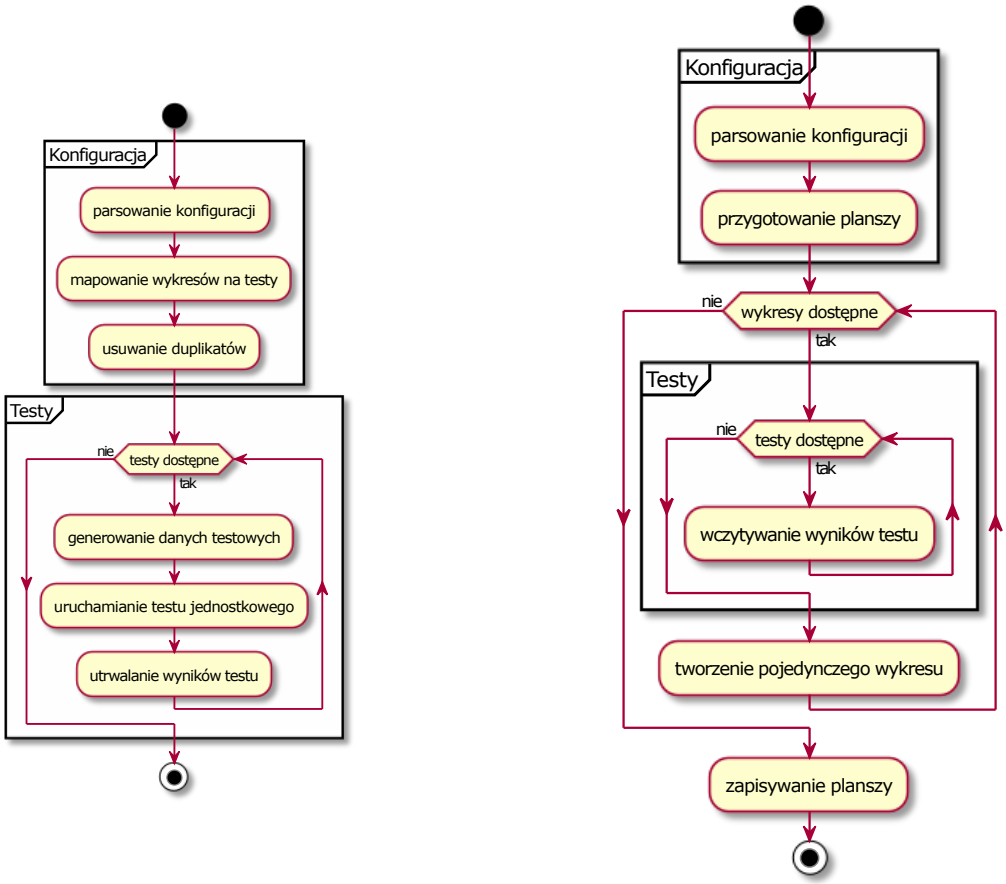
TODO: opis diagramu



Rysunek 5.1: Diagram klas silnika testującego

5.3.2 Diagram aktywności

TODO: opis diagramu



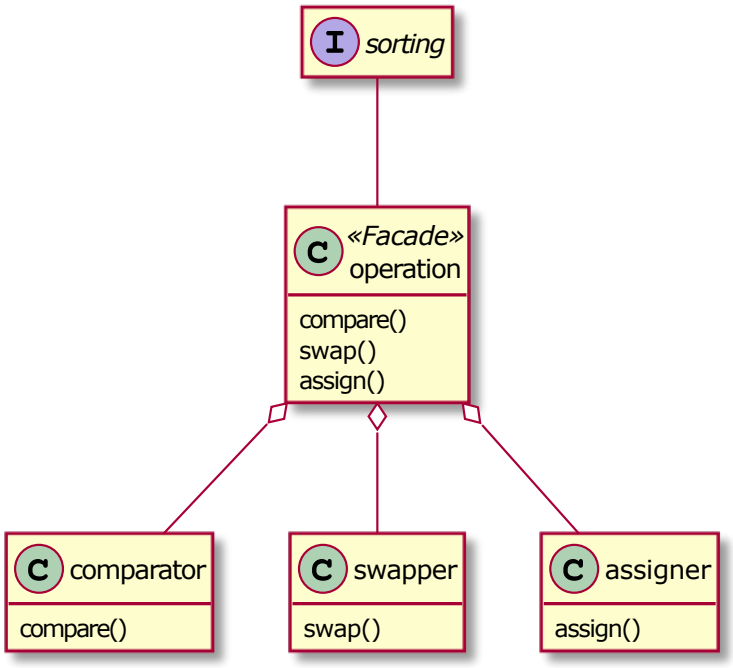
(a) silnik testujący (b) silnik graficzny

Rysunek 5.2: Diagram aktywności projektowanego systemu

5.4 Wzorce projektowe

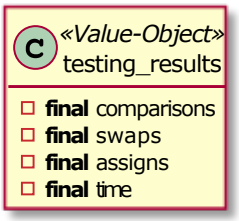
5.4.1 Fasada

W trakcie działania algorytm sortujący wykonuje wiele operacji atomowych, takich jak porównywanie elementów, zamiana elementów miejscami oraz operacje przypisania. Aby uniknąć nadmiaru odpowiedzialności dla klas sortujących zastosowano obiekt pośredniczący **operation** będący równocześnie **fasadą**. Fasada zapewnia jednolity interfejs dla wszystkich operacji atomowych oraz przekierowuje ich działanie do obiektów bezpośrednio odpowiedzialnych za ich wykonanie.



5.4.2 Obiekt-Wartość

Proces testowania algorytmu składa się z wielu iteracji. Każdy z atomowych testów wchodzących w skład iteracji powinien być całkowicie niezależny i odseparowana od innych testów. Aby to zapewnić, dane pochodzące z osobnych testów są przekazywane za pomocą **obiektów-wartości**. Pola w takim obiekcie po inicjalizacji stają się niemodyfikowalne. Użytkownik może jedynie odczytać ich wartość, bez możliwości ich modyfikacji. **Obiekt-wartość** jest gwarancją, że wyniki pochodzące z testu są rzetelne oraz nie zostały zmodyfikowane w trakcie przepływu danych pomiędzy procesami.



5.4.3 Strategia

5.4.4 Budowniczy

Podsumowanie

Podsumowanie wyników testowania algorytmów. Wnioski z analizy algorytmów hybrydowych.

Bibliografia

Słownik pojęć

A.1 Notacja O())

A.2 Algorytm działający w miejscu

A.3 Algorytm stabilny

Algorytm nie zamienia kolejnością elementów o tej samej wartości.

Środowisko uruchomieniowe aplikacji

Platforma uruchomieniowa aplikacji - Windows. Wykorzystane języki programowania - C++, Python.

B.1 Zmienne środowiskowe

Opis zmiennych środowiskowych TEST-DIRECTORY, CONFIG-DIRECTORY, PLOT-DIRECTORY.

B.2 Biblioteki zewnętrzne

Biblioteki w C++ oraz Pythonie potrzebne do uruchomienia aplikacji wraz z numerami wersji.

B.3 Instalowanie aplikacji

Uruchamianie skryptu zaciągającego potrzebne zależności. Uruchamianie pliku makefile kompilującego i instalującego aplikację. Cykl pracy programu - tworzenie konfiguracji, testowanie silnikiem testującym, wizualizacja wyników przy użycia silnika graficznego.

B.4 Przykładowy plik konfiguracyjny

Plik konfiguracyjny w jsonie. Omówienie pliku, na początku są dane współdzielone przez wszystkie testy. Potem plik zawiera listę elementów typu plot, czyli listę osobnych testów.