VILNIAUS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS
INFORMATIKOS INSTITUTAS
INFORMATIKOS KATEDRA

Kursinis darbas

Rikiavimo tobulinimas genetiniais algoritmais

(Improving sorting with genetic algorithms)

Atliko: 3 kurso 2 grupės studentas
Deividas Zaleskis (parašas)

Darbo vadovas:
Irmantas Radavičius (parašas)

Vilnius
2021

Turinys

Įvadas 2
1. Šelo rikiavimo algoritmas 5
1.1. Šelo rikiavimo algoritmo efektyvumas 5
1.1.1. Šelo rikiavimo algoritmo asimptotinė analizė 5
1.1.2. Šelo rikiavimo algoritmo efektyvumas priklausomai nuo duomenų dydžio 5
1.1.3. Šelo rikiavimo algoritmo efektyvumas priklausomai nuo duomenų specifikos 5
1.2. Šelo rikiavimo algoritmo versijos 6
1.2.1. Vadovėlinis Šelo rikiavimo algoritmas 6
1.2.2. Patobulintas Šelo rikiavimo algoritmas 6
2. Genetiniai algoritmai 8
2.1. Galima paprasto genetinio algoritmo implementacija 9
3. Tarpų sekų efektyvumo kriterijų nustatymas 10
4. Eksperimentų vykdymo aplinkos paruošimas 11
4.1. Techninės detalės 11
4.2. Pasiruošimas matavimų atlikimui 11
4.2.1. Operacijas skaičiuojantis vadovėlinis Šelo algoritmas 11
4.2.2. Operacijas skaičiuojantis patobulintas Šelo algoritmas 11
4.2.3. Veikimo laiko matavimas 11
5. Tarpų sekų generavimas 12
5.1. Trumpų tarpų sekų generavimas 12
5.2. Vidutinio ilgio tarpų sekų generavimas 12
6. Tarpų sekų efektyvumo įvertinimas 14
6.1. Pasiruošimas tarpų sekų vertinimui 14
6.2. Trumpų tarpų sekų efektyvumo įvertinimas 14
6.2.1. Trumpų tarpų sekų efektyvumo įvertinimas naudojant vadovėlinį Šelo algoritmą . 14
6.2.2. Trumpų tarpų sekų efektyvumo įvertinimas naudojant patobulintą Šelo algoritmą 16
6.3. Vidutinio ilgio tarpų sekų efektyvumo įvertinimas 16
Išvados 18
Literatūra 19
Priedas Nr.1
Priedas Nr.2

Įvadas

Viena pagrindinių informatikos sąvokų yra algoritmas. Formaliai algoritmą galima apibūdinti kaip baigtinę seką instrukcijų, nurodančių kaip rasti nagrinėjamo uždavinio sprendinį. Algoritmo koncepcija egzistuoja nuo senovės laikų [Knu72], tačiau atsiradus kompiuteriams, tapo įmanoma algoritmų vykdymą automatizuoti, paverčiant juos mašininiu kodu suprantamu kompiuteriams [WWG51]. Taip informatikos mokslas nuo teorinių šaknų [Tur37] įgavo ir taikomąją pusę. Beveik visus algoritmus galima suskirstyti į dvi klases: kombinatorinius algoritmus ir skaitinius algoritmus. Skaitiniai algoritmai sprendžia tolydžius uždavinius: optimizuoti realaus argumento funkciją, išspręsti tiesinių lygčių sistemą su realiais koeficientais, etc. Kombinatoriniai algoritmai sprendžia diskrečius uždavinius ir operuoja diskrečiais objektais: skaičiais, sąrašais, grafais, etc. Vienas žinomiausių diskretaus uždavinio pavyzdžių yra duomenų rikiavimas.

Duomenų rikiavimas yra vienas pamatinių informatikos uždavinių. Matematiškai jis formuluojamas taip: duotai baigtinei palyginamų elementų sekai $S = (s_1, s_2, ..., s_n)$ pateikti tokį kėlinį, kad pradinės sekos elementai būtų išdėstyti didėjančia (mažėjančia) tvarka [RB13]. Rikiavimo uždavinys yra aktualus nuo pat kompiuterių atsiradimo ir buvo laikomas vienu pagrindinių diskrečių uždavinių, kuriuos turėtų gebėti spręsti kompiuteris [Knu70]. Rikiavimo uždavinio sprendimas dažnai padeda pagrindą efektyviam kito uždavinio sprendimui, pavyzdžiui, atliekant paiešką sąraše, galima taikyti dvejetainės paieškos algoritmą tik tada, kai sąrašas yra išrikiuotas. Kadangi rikiavimo uždavinys yra fundamentalus, jam spręsti egzistuoja labai skirtingų algoritmų.

Rikiavimo algoritmų yra įvairių: paremtų palyginimu (elementų tvarką nustato naudojant palyginimo operatorius), stabilių (nekeičia lygių elementų tvarkos), nenaudojančių papildomos atminties (atminties sudėtingumas yra $O(1)$), etc. Asimptotiškai optimalūs palyginimu paremti algoritmai blogiausiu atveju turi $O(n \, log \, n)$ laiko sudėtingumą, o ne palyginimu paremti algoritmai gali veikti dar greičiau, tačiau nėra tokie universalūs, kadangi rikiuojama remiantis duomenų specifika. Tiesa, rikiuojant remtis vien algoritmo asimptotika nepakanka: rikiavimas įterpimu (angl. insertion sort) blogiausiu atveju turi $O(n^2)$ laiko sudėtingumą [BFM06], tačiau mažesnius elementų kiekius rikiuoja daug greičiau, nei asimptotiškai optimalūs algoritmai, pavyzdžiui, rikiavimas krūva (angl. heapsort) [For64]. Todėl pastaruoju metu plačiai naudojami hibridiniai rikiavimo algoritmai, kurie sujungia keletą rikiavimo algoritmų į vieną ir panaudoja jų geriausias savybes. Nepaisant įvairovės ir naujų algoritmų gausos, klasikiniai rikiavimo algoritmai išlieka aktualūs.

Šelo rikiavimo algoritmas (angl. Shellsort, toliau - Šelo algoritmas) [She59] yra paremtas palyginimu, nenaudojantis papildomos atminties ir nestabilus. Šelo algoritmą galima laikyti rikiavimo įterpimu modifikacija, kuri lygina ne gretimus, o toliau vienas nuo kito esančius elementus, taip paspartindama jų perkėlimą į galutinę poziciją. Pagrindinė algoritmo idėja - išskaidyti rikiuojamą seką S į posekius $S_1, S_2, ..., S_n$, kur kiekvienas posekis $S_i = (s_i, s_{i+h}, s_{i+2h}, ...)$ yra sekos S elementai, kurių pozicija skiriasi h. Išrikiavus visus sekos S posekius S_i su tarpu h, seka tampa h-išrikiuota. Remiantis tuo, jog sekai S esant h-išrikiuota ir ją k-išrikiavus, ji lieka h-išrikiuota [GK72], galima kiekvieną algoritmo iteraciją mažinti tarpą, taip vis didinant sekos S išrikiuotumą.

Pritaikant šias idėjas ir rikiavimui naudojant mažėjančią tarpų seką su paskutiniu nariu 1, kuris garantuoja rikiavimą įterpimu paskutinėje iteracijoje, galima užtikrinti, jog algoritmo darbo pabaigoje seka S bus pilnai išrikiuota. Įvertinant Šelo algoritmo idėjas, nesunku pastebėti tarpų sekų įtaką jo veikimui.

Šelo algoritmo efektyvumas tiesiogiai priklauso nuo pasirinktos tarpų sekos. Weiss atlikto tyrimo [Wei91] rezultatai rodo, jog su Sedgewick pasiūlyta seka šis algoritmas veikia beveik dvigubai greičiau nei Šelo pradinis variantas, kai $n = 1000000$. Yra įrodyta, kad Šelo algoritmo laiko sudėtingumo blogiausiu atveju apatinė riba yra $\Omega(\frac{n\,log^2\,n}{log\,log\,n^2})$ [PPS92], taigi jis nėra asimptotiškai optimalus. Tiesa, kol kas nėra rasta seka, su kuria Šelo algoritmas pasiektų šią apatinę ribą. Kiek žinoma autoriui, asimptotiškai geriausia tarpų seka yra rasta Pratt, kuri yra formos 2^p3^p ir turi $\Theta(n\,log^2\,n)$ laiko sudėtingumą [Pra72], tačiau praktikoje ji veikia lėčiau už Ciura [Ciu01] ar Tokuda [Tok92] pasiūlytas sekas. Daugelio praktikoje efektyvių sekų asimptotinis sudėtingumas laiko atžvilgiu lieka atvira problema, nes jos yra rastos eksperimentiškai. Vienas iš metodų, kuriuos galima taikyti efektyvių tarpų sekų radimui, yra genetinis algoritmas.

Genetinis algoritmas (GA) yra metodas rasti euristikas, paremtas biologijos žiniomis apie natūralios atrankos procesą. Kartu su genetiniu programavimu, evoliuciniais algoritmais ir kitais metodais, genetiniai algoritmai sudaro evoliucinių skaičiavimų šeimą. Visi šios šeimos atstovai yra paremti pradinės populiacijos generavimu ir iteraciniu populiacijos atnaujinimu naudojant biologijos įkvėptas strategijas. J.H. Holland, GA pradininkas, savo knygoje [Hol92] apibrėžė genetinio algoritmo sąvoką ir su ja glaudžiai susijusias chromosomų (potencialių uždavinio sprendinių, išreikštų genų rinkiniu), bei rekombinacijos (tėvinių chromosomų genų perdavimo palikuonims), atrankos (tinkamiausių chromosomų atrinkimo) ir mutacijos (savaiminio chromosomos genų kitimo) operatorių koncepcijas. Genetinių algoritmų veikimo strategija pagrįsta pradinės chromosomų populiacijos evoliucija, kiekvienos naujos chromosomų kartos gavimui naudojant rekombinacijos, atrankos ir mutacijos operatorius. Toliau bus aptariamos genetinių algoritmų taikymo galimybės.

Genetiniai algoritmai taikomi sprendžiant įvairius paieškos ir optimizavimo uždavinius, kuomet nesunku nustatyti, ar sprendinys tinkamas, tačiau tinkamo sprendinio radimas reikalauja daug resursų ar net pilno perrinkimo. Tokiu atveju apytikslio sprendinio radimas (euristika) gali būti daug patrauklesnis sprendimo būdas, kadangi tikslaus sprendinio radimas dažnai yra NP-sunkus uždavinys. Todėl GA yra pritaikomi sudarant grafikus ir tvarkaraščius, sprendžiant globalaus optimizavimo uždavinius ir net projektuojant NASA mikrosatelitų antenas [HGL+06]. Nesunku pastebėti, jog efektyvių Šelo algoritmo tarpų sekų radimas yra sunkus uždavinys atliekamų skaičiavimų prasme, tikėtinai reikalaujantis pilno potencialių sprendinių perrinkimo, tad šio uždavinio sprendimui taikyti GA yra prasminga. Kiek žinoma autoriui, kol kas yra buvę du bandymai taikyti genetinius algoritmus efektyvių Šelo algoritmo tarpų sekų radimui [RBH+02; SY99]. Abiejuose darbuose teigiama, jog genetiniais algoritmais gautos tarpų sekos veikia greičiau už Sedgewick seką, kuri literatūroje laikoma viena efektyviausių.

Darbo **tikslas**: pritaikyti genetinius algoritmus Šelo algoritmo tarpų sekoms generuoti.
Darbo uždaviniai:

- Atlikti Šelo rikiavimo algoritmo literatūros analizę.
- Nustatyti kriterijus tarpų sekų efektyvumui įvertinti.
- Paruošti aplinką eksperimentų vykdymui.
- Naudojant genetinius algoritmus sugeneruoti tarpų sekas.
- Atliekant eksperimentus įvertinti sugeneruotų ir pateiktų literatūroje tarpų sekų efektyvumą.

Šis darbas sudarytas iš 6 skyrių. Pirmame skyriuje atliekama Šelo rikiavimo algoritmo literatūros analizė. Antrame skyriuje apžvelgiama genetinio algoritmo sąvoka. Trečiame skyriuje nustatomi kriterijai tarpų sekų efektyvumui įvertinti. Ketvirtame skyriuje paruošiama eksperimentų vykdymo aplinka. Penktame skyriuje generuojamos tarpų sekos, naudojant genetinius algoritmus. Šeštame skyriuje atliekant eksperimentus įvertinamas sugeneruotų ir pateiktų literatūroje tarpų sekų efektyvumas.

1. Šelo rikiavimo algoritmas

1.1. Šelo rikiavimo algoritmo efektyvumas

1.1.1. Šelo rikiavimo algoritmo asimptotinė analizė

1.1.2. Šelo rikiavimo algoritmo efektyvumas priklausomai nuo duomenų dydžio

Renkantis kokį algoritmą naudoti, labai svarbu įvertinti tikėtiną duomenų dydį. Šelo rikiavimo algoritmas yra efektyviausias, kai duomenų dydis yra ganėtinai mažas [Ciu01]. Kaip teigiama [SY99], šis algoritmas veikia geriausiai, kai $N \simeq 106$. Tokiu atveju, jis lenkia net ir vienu greičiausių laikomą greitojo rikiavimo algoritmą. Todėl Šelo rikiavimo algoritmas dažnai naudojamas hibridiniuose rikiavimo algoritmuose, kai pasiekus tam tikrą rekursijos lygį rikiuojamų duomenų dalis tampa pakankamai maža ir tolimesnė rekursija asimptotiškai optimaliu algoritmu nebeturi prasmės, kadangi jos tęsimas reikalautų per daug mašinos resursų. Tokių šio algoritmo naudojimo pavyzdžių galima rasti Go programavimo kalbos standartinėje bibliotekoje [Aut09] bei bzip2 failų glaudinimo programoje [Sew10]. Tiesa, net ir rikiuojant didesnius duomenų dydžius, Šelo algoritmas nėra labai lėtas [Ciu01]. Todėl jis yra vertingas įrankis įgyvendinant operacinės sistemos branduolį ar programuojant įterptinėms sistemoms, kadangi dėl ribotų atminties išteklių asimptotiškai optimalūs algoritmai, kurie dažniausia yra rekursyvūs ir naudoja daugiau atminties, tokiais atvejais netinka. Taigi, nors Šelo rikiavimo algoritmas nėra vienas greičiausių kai uždavinys didelis, yra scenarijų, kuriais šį algoritmą rinktis tikrai verta.

1.1.3. Šelo rikiavimo algoritmo efektyvumas priklausomai nuo duomenų specifikos

Viena palankiausių Šelo algoritmo savybių yra jo adaptyvumas. Adaptyvumas rikiavimo algoritmų kontekste reiškia, jog algoritmas atlieka mažiau operacijų, jei duomenys dalinai išrikiuoti. Adaptyvumą Šelo algoritmas paveldi iš rikiavimo įterpimu, nes yra jo optimizacija. Verta pastebėti, jog Šelo algoritmas būtent šia savybe ir remiasi rikiavimo įterpimu optimizavimui: kadangi rikiavimas įterpimu veikia labai greitai, kai duomenys išrikiuoti (tokiu atveju sudėtingumas laiko atžvilgiu yra $O(n)$), duomenis pradžioje h-išrikiavus, paskutine iteracija taikomas rikiavimas įterpimu veikia daug greičiau. Moksliniuose tyrimuose ši savybė retai turi įtakos, kadangi dažniausia siekiama ištirti algoritmo veikimą, kai duomenys nėra tvarkingai išdėstyti ir tipiškai renkamasi atsitiktinai generuoti pradinius duomenis šiam tikslui pasiekti. Savaime aišku, jog praktikoje retai pavyks sutikti visiškai atsitiktinai išsidėsčiusius duomenis, neturinčius nei vieno išrikiuoto posekio ilgesnio nei 1. Kaip praktinį šios idėjos taikymo pavyzdį galima pateikti vieną populiariausių šiuo metu naudojamų algoritmų, Timo rikiavimo algoritmą (angl. Timsort) [AJN+18]. Timo rikiavimo algoritmas remiasi tuo, jog realiame pasaulyje sutinkami duomenys labai dažnai turi išrikiuotų posekių, o juos aptikus rikiavimą galima atlikti greičiau, kadangi šių posekių rikiuoti nebereikia. Kaip parodo [CK80], Šelo algoritmas veikia net keletą kartų greičiau, kai duomenys yra pilnai išrikiuoti ir veikimo sparta nusileidžia tik rikiavimui įterpimu su visais duomenų dydžiais. Kai duomenų

dydis nedidelis (iki 200 elementų) ir 20% elementų nėra teisingose pozicijose, Šelo algoritmas veikimo sparta nusileidžia tik greitojo rikiavimo algoritmui. Palankūs rezultatai pateikiami ir [FG15] - su dalinai išrikiuotais duomenimis Šelo algoritmas veikia apytiksliai 1.5 karto greičiau, o su išrikiuotais duomenimis pasiekiamas apie 3.8 karto mažesnis veikimo greitis. Taigi, rikiuojant dalinai išrikiuotus duomenis Šelo algoritmas yra labai efektyvus.

1.2. Šelo rikiavimo algoritmo versijos

1.2.1. Vadovėlinis Šelo rikiavimo algoritmas

Kaip ir daugelis algoritmų, Šelo rikiavimo algoritmas turi keletą galimų implementacijų. Žinomiausia iš jų, be abejo, yra vadovėlinė šio algoritmo versija, daugeliui pasaulio informatikos studentų žinoma iš duomenų struktūrų ir algoritmų kurso. Jos pseudokodas pateikiamas žemiau.

Algorithm 1 Vadovėlinis Šelo rikiavimo algoritmas

1: **foreach** gap **in** H **do**
2: **for** $i \leftarrow gap + 1$ **to** N **do**
3: $j \leftarrow i$
4: $temp \leftarrow S[i]$
5: **while** $j > gap$ **and** $S[j - gap] > S[j]$ **do**
6: $S[j] \leftarrow S[j - gap]$
7: $j \leftarrow j - gap$
8: **end while**
9: $S[j] \leftarrow temp$
10: **end for**
11: **end for**

1.2.2. Patobulintas Šelo rikiavimo algoritmas

Verta pastebėti, jog vadovėlinė Šelo algoritmo implementacija nėra naši [RB13]. Vykdant vidinį vadovėlinio Šelo algoritmo ciklą (1.5 - 1.8 eilutės), 1.5 eilutėje yra tikrinama, ar $S[j]$ jau yra tinkamoje pozicijoje. Jei $S[j]$ jau yra tinkamoje pozicijoje, vidinis ciklas nėra vykdomas ir jokių elementų pozicijos nėra keičiamos. Tačiau 1.4 ir 1.9 eilutėse vis tiek yra vykdomi du priskyrimai, kurie šiuo atveju nueina veltui (o taip nutinka pakankamai dažnai). Siekiant tai ištaisyti, [RB13] pateikė patobulintą Šelo algoritmo versiją, kuri prieš vykdydama bet kokias kitas instrukcijas patikrina, ar elementas $S[j]$ jau yra tinkamoje pozicijoje. Patobulinto Šelo algoritmo pseudokodas pateikiamas žemiau.

Algorithm 2 Patobulintas Šelo rikiavimo algoritmas

```
1: foreach gap in H do
2:     for i ← gap + 1 to N do
3:         if S[i − gap] > S[i] then
4:             j ← i
5:             temp ← S[i]
6:             repeat
7:                 S[j] ← S[j − gap]
8:                 j ← j − gap
9:             until j ≤ gap or S[j − gap] ≤ S[j]
10:            S[j] ← temp
11:        end if
12:    end for
13: end for
```

2. Genetiniai algoritmai

Prieš pradedant generuoti sekas, reikalinga plačiau apžvelgti genetinio algoritmo sąvoką. Paprasčiausias genetinis algoritmas susideda iš chromosomų populiacijos bei atrankos, mutacijos ir rekombinacijos operatorių. Projektuojant genetinį algoritmą tam tikro uždavinio sprendimui, svarbu tinkamai pasirinkti, kaip kompiuteriu modeliuoti galimus sprendinius (chromosomas). Chromosomos kompiuterio atmintyje standartiškai išreiškiamos bitų eilutėmis, kadangi tai palengvina tiek mutaciją (pakanka apversti kurio nors atsitiktinio bito reikšmę), tiek rekombinaciją (pakanka perkopijuoti pasirinktus tėvinių chromosomų bitus į vaikinę chromosomą). Tiesa, tai nėra vienintelis įmanomas būdas, ir kai kurių uždavinių sprendiniai modeliuojami pvz. grafu ar simbolių eilute. Chromosomų rinkinys, literatūroje dažnai vadinamas populiacija, atspindi uždavinio sprendinių aibę, kuri kinta kiekvieną algoritmo iteraciją. Sprendinio kokybę įvardijame kaip jo tinkamumą, kuris apibrėžiamas tinkamumo funkcijos reikšme, pateikus sprendinį kaip parametrą. Nesunku pastebėti, jog tinkamumo funkcija tėra tikslo funkcijos specializacija, kuri naudojama chromosomų vertinimui. Tinkamumo funkcija yra viena svarbiausių genetinio algoritmo dalių, kadangi kai ji netinkamai parinkta, algoritmas nekonverguos į tinkamą sprendinį arba užtruks labai ilgai. Genetinis algoritmas vykdymo metu iteratyviai atnaujina esamą populiaciją, kurdamas naujas kartas taikant atrankos, rekombinacijos ir mutacijos operatorius. Atrankos operatorius grąžina tinkamiausius populiacijos individus, kuriems yra leidžiama susilaukti palikuonių taikant rekombinacijos operatorių. Rekombinacijos operatorius veikia iš dviejų tėvinių chromosomų sukurdamas naują vaikinę chromosomą, kas dažniausiai pasiekiama tam tikru būdu perkopijuojant tėvų genų atkarpas į vaikinę chromosomą. Rekombinacijos strategijų yra įvairių, tačiau ne kiekvienam uždaviniui visos jos tinka, kadangi kai kuriais atvejais netinkamai parinkta rekombinacijos strategija pagamina neatitinkančią uždavinio apribojimų vaikinę chromosomą. Pavyzdžiui, jei modeliuojama chromosoma yra sąrašas, turintis susidėti iš tam tikrų elementų, neįmanoma garantuoti, jog atsitiktinai perkopijavus tėvinių chromosomų genus į vaikinę chromosomą šis apribojimas bus išlaikytas. Galiausiai tam tikrai populiacijos daliai yra pritaikomas mutacijos operatorius. Jo veikimo principas yra gana paprastas: pasirinktos chromosomos vienas ar keli genai yra modifikuojami, nežymiai pakeičiant jų reikšmes ar sukeičiant kelių genų reikšmes vietomis. Mutacijos operatorius praplečia vykdomos paieškos erdvę, o tai labai svarbu, kadangi kitaip algoritmas gali konverguoti į lokaliojo minimumo taškus, taip ir nepasiekdamas globaliojo minimumo. Atsižvelgiant į tai, yra laikoma, jog mutacijos operatorius yra kertinė genetinio algoritmo dalis, kuri palaiko genetinę individų įvairovę ir padeda rasti tinkamiausius sprendinius. Jei algoritmas suprojektuotas tinkamai, dažnu atveju vidutinis populiacijos tinkamumas gerės kiekvieną iteraciją. Kad algoritmas neveiktų amžinai, yra pasirenkama tam tikra sustojimo sąlyga, pvz. pasiektas maksimalus kartų skaičius, vidutinio tinkamumo pokytis tarp paskutinių dviejų populiacijos kartų pakankamai mažas.

2.1. Galima paprasto genetinio algoritmo implementacija

Apjungiant visas aukščiau aptartas genetinio algoritmo dalis, galima suformuluoti paprasto genetinio algoritmo pseudokodą.

Algorithm 3 Paprastas GA

Let $current\_population$ be a random population of chromosomes.
Let $mutation\_rate$ be a real value between 0 and 1.
Let $recombination\_fraction$ be a real value between 0 and 1.
repeat
    Apply the fitness function to each chromosome in $current\_population$.
    Sort the chromosomes in $current\_population$ based on their fitness.
    Let $new\_population$ be an empty set.
    repeat
        Let $(parent1, parent2)$ be a pair of parent chromosomes selected randomly from the first $[N * recombination\_fraction]$ elements of $current\_population$.
        Let $child\_chromosome$ be the result of applying the recombination operator to $parent1$ and $parent2$.
        Let $rand$ be a random real value between 0 and 1.
        if $rand < mutation\_rate$ **then**
            Apply the mutation operator to $child\_chromosome$
        end if
        Add $child\_chromosome$ to $new\_population$
    until N offspring have been created.
    Assign $new\_population$ to $current\_population$.
until termination criteria are satisfied.

3. Tarpų sekų efektyvumo kriterijų nustatymas

Šelo algoritmo tarpų sekų efektyvumo įvertinimas nėra trivialus. Rikiavimo algoritmai dažniausiai yra vertinami pagal atliekamų priskyrimų skaičių. Skaičiuojant algoritmo atliekamus priskyrimus, gana paprasta jais išreikšti inversijų skaičių. Daugelyje algoritmų priskyrimų skaičius uždaviniui augant greitai artėja prie palyginimų skaičiaus, tad tokiu metodu gautas įvertis būna pakankamai tikslus. Šelo algoritmo atveju, vien priskyrimų skaičius nėra pakankamai tikslus kriterijus tarpų sekų efektyvumui įvertinti, kadangi remiantis tik juo, gautas įvertis neatspindi praktinio efektyvumo. Kaip parodo [Ciu01], šiame algoritme dominuojanti operacija yra palyginimas. Tad galima daryti išvadą, jog atliekamų palyginimų skaičius yra tinkamesnis kriterijus efektyvumui įvertinti.

Matuojant rikiavimo algoritmo efektyvumą tik naudojant sveikaskaitinius pradinius duomenis, gauti rezultatai gali būti netikslūs, kadangi šių operacijų sparta priklauso nuo rikiuojamų duomenų tipo. Rikiuojant simbolių eilutes, priskyrimas atliekamas naudojant rodykles, kas yra $O(1)$ operacija, tačiau palyginimas yra $O(n)$ blogiausiu atveju. Ir atvirkščiai, rikiuojant įrašus kurie saugomi steke, priskyrimas reikalauja perkopijuoti visą įrašą, o palyginimas gali būti atliekamas naudojant tam tikrą raktą. Todėl apsiriboti vien palyginimų ar priskyrimų skaičiumi nepakanka, kadangi tiksliausia praktinio algoritmo veikimo laiko aproksimacija (tai, kam ir skaičiuojamos operacijos), bus gauta tik įvertinant abu šiuos rodiklius.

Algoritmo veikimo laikas, nors ir priklausomas nuo platformos, kurioje vykdomas tyrimas, detalių, taip pat gali duoti tinkamų įžvalgų įvertinant praktinį efektyvumą. Kadangi algoritmo atliekamos operacijos skaičiuojamos tam, jog gauti praktinio veikimo laiko aproksimaciją, tai realus veikimo laikas yra konkretus įvertis, leidžiantis praktiškai įvertinti duotos sekos efektyvumą. Taigi, įvertinant tarpų sekų efektyvumą bus remiamasi visomis atliekamomis operacijomis bei realiais veikimo laikais.

4. Eksperimentų vykdymo aplinkos paruošimas

4.1. Techninės detalės

Tyrimui buvo naudotas kompiuteris su 2.70 GHz Intel(R) Core(TM) i7-10850H procesoriumi, 32 GB operatyviosios atminties ir Windows 10 operacine sistema. Tyrimas buvo įgyvendintas C++ kalba su GNU g++ 8.1.0 kompiliatoriumi. Genetinio algoritmo implementacijai buvo pasirinkta OpenGA biblioteka [MAM+17] dėl suteikiamos laisvės pasirinkti, kaip įgyvendinti genetinius operatorius bei modernių kalbos konstruktų ir lygiagretaus vykdymo palaikymo.

4.2. Pasiruošimas matavimų atlikimui

4.2.1. Operacijas skaičiuojantis vadovėlinis Šelo algoritmas

Siekant išmatuoti vadovėlinio Šelo algoritmo atliekamų operacijų skaičių, buvo parengtas **??** algoritmas, kuris kaip rezultatą grąžina atliktus palyginimus bei priskyrimus.

4.2.2. Operacijas skaičiuojantis patobulintas Šelo algoritmas

Siekant išmatuoti patobulinto Šelo algoritmo atliekamų operacijų skaičių, buvo parengtas **??** algoritmas, kuris kaip rezultatą grąžina atliktus palyginimus bei priskyrimus.

4.2.3. Veikimo laiko matavimas

Verta pastebėti, jog matuoti veikimo laiką naudojant **??** ir **??** algoritmus nėra tinkama, kadangi jie skaičiuodami atliekamas operacijas vykdo papildomus žingsnius, o tai gali iškreipti gautus rezultatus. Atsižvelgiant į tai, veikimo laiko matavimui naudoti 1 ir 2 algoritmai.

5. Tarpų sekų generavimas

5.1. Trumpų tarpų sekų generavimas

Pirmame etape buvo generuojamos tarpų sekos, kurios efektyvios kai $N = 1000$. Remiantis [SY99], trumpos sekos chromosoma buvo modeliuojama kaip septynių sveikų skaičių masyvas. Trumpų sekų generavimui buvo paruoštas vienkriterinis genetinis algoritmas. Chromosomos buvo vertinamos jas naudojant 20-ties atsitiktinai sugeneruotų sveikų skaičių masyvų rikiavimui vadovėline Šelo rikiavimo algoritmo versija ir skaičiuojant atliktas palyginimo operacijas. Chromosomos tinkamumo funkcija buvo apibrėžta kaip atliktų palyginimų skaičiaus aritmetinis vidurkis. Rekombinacijos operatorius buvo įgyvendintas tolygia strategija, kur abiejų tėvų genai turi vienodą tikimybę būti perduoti vaikinei chromosomai. Mutacijos operatorius buvo įgyvendintas su $\frac{1}{5}$ tikimybe keičiant tam tikrą chromosomą, pridedant prie atsitiktinai pasirinkto geno reikšmės skaičių $(rand() - rand()) * 25 * \frac{2}{10*\sqrt{kartos\_numeris}}$ su apribojimu, jog pakeistas genas turi priklausyti intervalui $[1, 1000]$. Siekiant išvengti netinkamų sprendinių, kiekviena seka po rekombinacijos ar mutacijos operatorių taikymo buvo išrikiuojama ir užtikrinama, jog paskutinis sekos narys yra 1.

Algoritmas buvo vykdomas su atsitiktinai sugeneruota 10000 individų populiacija, kiekvieną iteraciją pritaikant rekombinacijos operatorių $\frac{1}{2}$ populiacijos. Tokiu būdu buvo sugeneruota ir atrinkta 10 sekų. Po to iš visų sugeneruotų sekų buvo atrinkta tinkamiausia, sugeneruotas sekas naudojant 10000 atsitiktinai sugeneruotų masyvų ilgio 1000 rikiavimui ir matuojant atliktų palyginimo operacijų aritmetinį vidurkį. Atlikus matavimus, buvo pasirinkta seka $855, 264, 86, 35, 12, 5, 1$.

5.2. Vidutinio ilgio tarpų sekų generavimas

Antrame etape buvo generuojamos tarpų sekos, kurios efektyvios kai $N = 100000$. Vidutinio ilgio sekų generavimui buvo pasitelktas modifikuotas genetinis algoritmas, naudotas generuojant trumpas tarpų sekas. Tarpų sekos chromosoma buvo modeliuojama kaip sveikų skaičių masyvas ilgio 12. Buvo atsisakyta chromosomų vertinimo naudojant 20 skirtingų masyvų, kadangi tai darė neigiamą įtaką genetinio algoritmo veikimo laikui. Vietoje to, kiekvienos sekos vertinimui buvo naudojami 5 atsitiktinai sugeneruoti masyvai ilgio 100000. Mutacijos operatorius buvo pakoreguotas ir su $\frac{1}{5}$ tikimybe keitė tam tikrą chromosomą, pridedant prie atsitiktinai pasirinkto geno reikšmės skaičių $(rand() - rand()) * 75 * \frac{2}{10*\sqrt{kartos\_numeris}}$.

Algoritmas buvo vykdomas su atsitiktinai sugeneruota 1000 individų populiacija, paskutinius septynis chromosomos genus inicializuojant elementais $855, 264, 86, 35, 12, 5, 1$, t.y. sekos, gautos pirmame etape, nariais. Siekiant išlaikyti tinkamiausius sprendinius, buvo pasitelktas elitizmas - geriausiai pasirodžiusios 75 chromosomos nekeistos patekdavo į kitą algoritmo iteraciją. Rekombinacijos operatorius buvo taikomas $\frac{1}{2}$ populiacijos kiekvieną algoritmo iteraciją. Tokiu būdu buvo sugeneruota ir atrinkta 10 sekų. Po to iš visų sugeneruotų sekų buvo atrinkta tinkamiausia, sugeneruotas sekas naudojant 1000 atsitiktinai sugeneruotų masyvų ilgio 100000 rikiavimui ir matuojant atliktų palyginimo operacijų aritmetinį vidurkį. Atlikus matavimus, buvo pasirinkta seka

$$45794, 17396, 7414, 3136, 1206, 561, 264, 86, 35, 12, 5, 1.$$

6. Tarpų sekų efektyvumo įvertinimas

6.1. Pasiruošimas tarpų sekų vertinimui

Iš anksto verta pastebėti, jog ne visos šiame skyriuje tiriamos tarpų sekos iš tiesų yra būtent tokio ilgio, koks pateikiamas šiame darbe. Tačiau būtina atsižvelgti į faktą, jog tai galėtų sukelti nelygias sąlygas atliekamuose matavimuose, kadangi seka, turinti elementų didesnių už maksimalų tiriamą N, atliktų nereikalingas operacijas, o per trumpa seka veiktų neefektyviai. Siekiant standartizuoti tarpų sekų ilgius, buvo pasirinkta kai kurias sekas sutrumpinti (ar pratęsti, jei tam yra žinoma rekursyvi formulė) taip, jog standartizuota seka būtų ilgiausia įmanoma seka, kurios visi elementai mažesni už maksimalų tiriamą N. Atliekamų matavimų skaičius buvo pasirinktas atsižvelgiant į tame etape tiriamo N dydį: jei N labai mažas, egzistuoja didelė tikimybė jog gautas vidutinis įvertis bus netikslus, matuojant, pavyzdžiui, tik 1000 kartų. Todėl tiriant trumpesnes sekas buvo stengtasi atlikti daugiau matavimų.

6.2. Trumpų tarpų sekų efektyvumo įvertinimas

Trumpų tarpų sekų efektyvumo tyrimui buvo pasirinktos šios tarpų sekos:

- Tokuda: 525,233,103,46,20,9,4,1
- Ciura: 701,301,132,57,23,10,4,1
- Simpson-Yachavaram: 893,219,83,36,13,4,1
- Sedgewick: 929,505,209,109,41,19,5,1
- Incerpi-Sedgewick: 861,336,112,48,21,7,3,1
- Trumpa seka: 855,264,86,35,12,5,1

Tokuda [Tok92] seka buvo pasirinkta atsižvelgiant į tai, jog praktikoje ji laikoma viena efektyviausių. Kiek žinoma autoriui, kai $N = 1000$ Ciura [Ciu01] seka atlieka mažiausiai palyginimų vidutiniu atveju, tad ją buvo būtina įtraukti. Simpson ir Yachavaram [SY99] seka buvo pasirinkta, kadangi pasak autorių, tai geriausiai veikianti jų rasta seka, kai $N = 1000$. Sedgewick [Sed86] ir Incerpi-Sedgewick [IS85] sekos pasirinktos dėl tvirto matematinio pagrindo (įrodytas $O(n^{\frac{4}{3}})$ sudėtingumas blogiausiu atveju) bei fakto, jog ilgą laiką šios sekos buvo laikomos vienomis efektyviausių.

6.2.1. Trumpų tarpų sekų efektyvumo įvertinimas naudojant vadovėlinį Šelo algoritmą

matavimai atlikti 50000 kartu.

1 lentelė. Vidutinis vadovėlinio Šelo algoritmo atliekamų palyginimų skaičius naudojant trumpas tarpų sekas

N	Tokuda	Ciura	S/Y	Sedgewick	I/S	TS
50	289.902	287.76	287.704	293.414	296.373	289.207
100	735.822	733.353	731.492	752.139	759.577	731.877
200	1812.59	1797.92	1798.5	1840.15	1870.94	1793.19
400	4315.6	4276.38	4284.94	4405.07	4464.59	4273.84
800	10061.3	9946.54	9987.18	10260.5	10528.2	9999.9
1000	13128.9	13044.4	13075.4	13395.4	13807.4	13054.4

2 lentelė. Vidutinis vadovėlinio Šelo algoritmo atliekamų priskyrimų skaičius naudojant trumpas tarpų sekas

N	Tokuda	Ciura	S/Y	Sedgewick	I/S	TS
50	482.372	473.174	453.477	447.686	487.444	456.583
100	1204.5	1186.33	1137.21	1130.61	1227.55	1135.35
200	2933.02	2862.17	2752.73	2761.21	2977.03	2745.81
400	6898.23	6733.29	6522.93	6629.27	6903.13	6444.15
800	15899.2	15487.9	14758.6	15354.8	15950	14745.8
1000	20707.8	20273.4	19200.7	20089.7	20841.4	19193.6

3 lentelė. Vidutinis vadovėlinio Šelo algoritmo veikimo laikas naudojant trumpas tarpų sekas

N	Tokuda	Ciura	S/Y	Sedgewick	I/S	TS
50	0.99916	1.02294	1.0962	1.00912	1.13534	0.8968
100	3.03872	2.87242	2.74936	2.9471	2.74198	3.09066
200	6.34908	6.24564	6.32294	6.00468	7.19932	6.37474
400	16.3897	15.0581	14.2942	14.2024	14.7652	14.9239
800	39.9326	37.5984	35.4103	38.5972	43.2178	33.2975
1000	45.7384	44.2994	42.0728	42.58	44.4423	44.0977

6.2.2. Trumpų tarpų sekų efektyvumo įvertinimas naudojant patobulintą Šelo algoritmą

4 lentelė. Vidutinis trumpų tarpų sekų atliekamų palyginimų skaičius

N	Matavimų sk.	Tokuda	Ciura	SY	Sedgewick	IS	GASTS
100	10000000	736.122	733.283	731.542	751.776	759.732	731.978
250	10000000	2401.49	2377.67	2384.14	2452.1	2475.83	2380.15
500	1000000	5665.78	5622.06	5630.74	5784.2	5895.15	5628.45
750	1000000	9315.87	9190.4	9230.42	9492.19	9719.5	9225.26
1000	1000000	13129.6	13043.9	13076.8	13394.2	13807	13052.5

5 lentelė. Vidutinis trumpų tarpų sekų atliekamų priskyrimų skaičius

N	Matavimų sk.	Tokuda	Ciura	SY	Sedgewick	IS	GASTS
100	10000000	790.912	797.569	811.157	853.562	822.576	814.139
250	10000000	2533.62	2571.78	2625.46	2694.74	2650.55	2670.51
500	1000000	5998.37	5994.61	6220.64	6361.64	6298.07	6275.24
750	1000000	9715.91	9815.62	10273.2	10265.8	10516.5	10330.8
1000	1000000	13864	13818.8	14509.9	14525.1	14867.6	14489.5

6 lentelė. Vidutinis trumpų tarpų sekų veikimo laikas

N	Matavimų sk.	Tokuda	Ciura	SY	Sedgewick	IS	GASTS
100	10000000	3.06033	2.99658	2.95647	2.8657	3.04825	3.02878
250	10000000	9.83692	9.64667	9.50202	9.4156	9.80572	9.66899
500	1000000	24.2449	23.5071	23.2876	22.6003	23.7591	22.0778
750	1000000	37.831	37.1193	35.9676	36.0782	37.484	36.4349
1000	1000000	53.2496	52.9506	51.282	52.0741	54.599	51.0197

6.3. Vidutinio ilgio tarpų sekų efektyvumo įvertinimas

Vidutinio ilgio tarpų sekų efektyvumo tyrimui buvo pasirinktos šios tarpų sekos:

- Tokuda: $68178,30301,13467,5985,2660,1182,525,233,103,46,20,9,4,1$
- Ciura: $90927,40412,17961,7983,3548,1577,701,301,132,57,23,10,4,1$
- Simpson-Yachavaram: $38291,22927,8992,3568,1488,893,219,83,36,13,4,1$
- Incerpi-Sedgewick: $86961,33936,13776,4592,1968,861,336,112,48,21,7,3,1$
- Sedgewick: $64769,36289,16001,8929,3905,2161,929,505,209,109,41,19,5,1$

- Roos-Bennet-Hannon-Zehner: $91433,72985,13229,5267,2585,877,155,149,131,23,8,1$
- GA sugeneruota vidutinio ilgio seka: $45794,17396,7414,3136,1206,561,264,86,35,12,5,1$

Išvados

Gavome tą ir aną, rekomenduojame šitai.

Literatūra

[AJN+18] Nicolas Auger, Vincent Jugé, Cyril Nicaud ir Carine Pivoteau. On the worst-case complexity of TimSort. *arXiv preprint arXiv:1805.08612*, 2018.

[Aut09] The Go Authors. Sort implementation in Go standard library. 2009. URL: https://golang.org/src/sort/sort.go (tikrinta 2021-05-24).

[BFM06] Michael A Bender, Martin Farach-Colton ir Miguel A Mosteiro. Insertion sort is O (n log n). *Theory of Computing Systems*, 39(3):391–397, 2006.

[Ciu01] Marcin Ciura. Best increments for the average case of shellsort. *International Symposium on Fundamentals of Computation Theory*, p.p. 106–117. Springer, 2001.

[CK80] Curtis R. Cook ir Do Jin Kim. Best Sorting Algorithm for Nearly Sorted Lists. *Commun. ACM*, 23(11):620–624, 1980-11. ISSN: 0001-0782. DOI: 10.1145/359024.359026. URL: https://doi.org/10.1145/359024.359026.

[FG15] Neetu Faujdar ir Satya Prakash Ghrera. Analysis and Testing of Sorting Algorithms on a Standard Dataset. *2015 Fifth International Conference on Communication Systems and Network Technologies*, p.p. 962–967, 2015. DOI: 10.1109/CSNT.2015.98.

[For64] G. E. Forsythe. Algorithms. *Commun. ACM*, 7(6):347–349, 1964-06. ISSN: 0001-0782. DOI: 10.1145/512274.512284. URL: https://doi.org/10.1145/512274.512284.

[GK72] David Gale ir Richard M. Karp. A phenomenon in the theory of sorting. *Journal of Computer and System Sciences*, 6(2):103–115, 1972. ISSN: 0022-0000. DOI: https://doi.org/10.1016/S0022-0000(72)80016-3. URL: https://www.sciencedirect.com/science/article/pii/S0022000072800163.

[HGL+06] Gregory Hornby, Al Globus, Derek Linden ir Jason Lohn. Automated antenna design with evolutionary algorithms. *Space 2006*, p. 7242. 2006.

[Hol92] John Henry Holland. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press, 1992.

[IS85] Janet Incerpi ir Robert Sedgewick. Improved upper bounds on Shellsort. *Journal of Computer and System Sciences*, 31(2):210–224, 1985.

[Knu70] Donald E. Knuth. Von Neumann's First Computer Program. *ACM Comput. Surv.*, 2(4):247–260, 1970-12. ISSN: 0360-0300. DOI: 10.1145/356580.356581. URL: https://doi.org/10.1145/356580.356581.

[Knu72] Donald E Knuth. Ancient babylonian algorithms. *Communications of the ACM*, 15(7):671–677, 1972.

[MAM+17] Arash Mohammadi, Houshyar Asadi, Shady Mohamed, Kyle Nelson ir Saeid Nahavandi. OpenGA, a C++ Genetic Algorithm Library. *Systems, Man, and Cybernetics (SMC), 2017 IEEE International Conference on*, p.p. 2051–2056. IEEE, 2017.

[PPS92] C. G. Plaxton, B. Poonen ir T. Suel. Improved lower bounds for Shellsort. *Proceedings., 33rd Annual Symposium on Foundations of Computer Science*, p.p. 226–235, 1992. DOI: 10.1109/SFCS.1992.267769.

[Pra72] Vaughan R Pratt. Shellsort and sorting networks. Tech. atask., STANFORD UNIV CA DEPT OF COMPUTER SCIENCE, 1972.

[RB13] Irmantas Radavičius ir Mykolas Baranauskas. An empirical study of the gap sequences for Shell sort. *Lietuvos matematikos rinkinys*, 54(A):61–66, 2013-12. DOI: 10.15388/LMR.A.2013.14. URL: https://www.journals.vu.lt/LMR/article/view/14899.

[RBH+02] Robert S Roos, Tiffany Bennett, Jennifer Hannon ir Elizabeth Zehner. A genetic algorithm for improved shellsort sequences. *Proceedings of the 4th Annual Conference on Genetic and Evolutionary Computation*, p.p. 694–694, 2002.

[Sed86] Robert Sedgewick. A new upper bound for Shellsort. *Journal of Algorithms*, 7(2):159–173, 1986. ISSN: 0196-6774. DOI: https://doi.org/10.1016/0196-6774(86)90001-5. URL: https://www.sciencedirect.com/science/article/pii/0196677486900015.

[Sew10] Julian Seward. Sort implementation in bzip2. 2010. URL: https://www.ncbi.nlm.nih.gov/IEB/ToolBox/CPP_DOC/lxr/source/src/util/compress/bzip2/blocksort.c#L519 (tikrinta 2021-05-24).

[She59] D. L. Shell. A High-Speed Sorting Procedure. *Commun. ACM*, 2(7):30–32, 1959-07. ISSN: 0001-0782. DOI: 10.1145/368370.368387. URL: https://doi.org/10.1145/368370.368387.

[SY99] Richard Simpson ir Shashidhar Yachavaram. Faster shellsort sequences: A genetic algorithm application. *Computers and Their Applications*, p.p. 384–387, 1999.

[Tok92] Naoyuki Tokuda. An Improved Shellsort. *Proceedings of the IFIP 12th World Computer Congress on Algorithms, Software, Architecture - Information Processing '92, Volume 1 - Volume I*, p.p. 449–457, NLD. North-Holland Publishing Co., 1992. ISBN: 044489747X.

[Tur37] Alan Mathison Turing. On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London mathematical society*, 2(1):230–265, 1937.

[Wei91] Mark Allen Weiss. Short Note: Empirical study of the expected running time of Shellsort. *The Computer Journal*, 34(1):88–91, 1991.

[WWG51] Maurice V Wilkes, David J Wheeler ir Stanley Gill. *The Preparation of Programs for an Electronic Digital Computer: With special reference to the EDSAC and the Use of a Library of Subroutines*. Addison-Wesley, 1951.

Priedas Nr. 1

Operacijas skaičiuojantis vadovėlinis Šelo algoritmas

Priedas Nr. 2

Operacijas skaičiuojantis patobulintas Šelo algoritmas