**Uvod**

Porast količine tekstualnih informacija u zadnja dva desetljeća potaknuo je razvoj struktura podataka koje će tu informaciju memorijski efikasno pohraniti, omogućujući pritom vremensku efikasnost pretrage i dohvata. Jedno od područja koje ima veliku korist od razvoja ovakvih struktura je bioinformatika. Radi same prirode bioinformatičkih problema, odnosno, obrade bioloških podataka (npr. analiza dugačkih sekvenci DNK), vrlo je pogodno nad takvim biološkim podacima izgraditi spomenute strukture kako bi se njihova analiza ubrzala.

Jedan od pristupa rješavanja ovog problema je upotreba indeksa pretraživanja po cijelom tekstu (eng. full-text search). Ove strukture omogućuju brzo i potpuno pretraživanje teksta nad kojim su izgrađene. Nakon izgradnje, pronalazak proizvoljnog tekstualnog uzorka (zajedno s njegovim brojem ponavljanja) vrlo je učinkovit. No, ovaj pristup, koji stavlja naglasak na brz pronalazak traženih uzoraka, često ne ispunjava memorijske zahtjeve. Budući da korisnost pretrage raste s količinom teksta, memorijska efikasnost vrlo je bitan faktor pri odabiru načina obrade i pohrane teksta kojega je potrebno analizirati.

Novije izvedbe indeksa dobiveni tekst prvo sažmu, te tek nakon toga, nad sažetim tekstom grade indekse. Jedan od takvih indeksa je FM-index čija je implementacija i zadatak ovoga projekta.

**Pomoćni algoritmi**

Wavelet stablo

Wavelet stablo (eng. wavelet tree) je podatkovna struktura koja tekstualni niz organizira u hijerarhiju nizova nula i jedinica - *bit-vektora*. Ova struktura omogućuje pronalazak broja pojavljivanja do nekog indeksa u nizu - *ranga* nekog znaka sa složenošću od O(log n), gdje je n veličina abecede danog tekstualnog niza.

Stablo se izgrađuje rekurzivno, razdijeljujući abecedu u parove podabeceda. Nakon izgradnje stabla, svaki list odgovara jednom znaku abecede, dok ostali unutarnji čvorovi označavaju da li pojedini znak teksta pripada prvoj ili drugoj podabecedi.

Ukoliko se *bit-vektori* koji izgrađuju stablo pohrane u RRR strukturu (koja je opisana u idućem poglavlju), moguće je da se memorijsko zauzeće smanji s obzirom na originalno stablo koje ne koristi RRR strukturu.

Konstruiranje wavelet stabla

Wavelet stablo pretvara dovedeni niz u balansirano binarno stablo *bit-vektora,* gdje se dio znakova, koji pripadaju prvoj polovici dane abecede, kodira kao 0, a drugi dio kao 1. Iako se na prvi pogled čini kako ovo unosi nejednoznačnost u dekodiranju sadržaja stabla, to ipak nije tako. Naime, prelaskom u slijedeću razinu, prva polovica abecede iz prethodne razine ponovno se dijeli na dva dijela te se njeni znakovi ponovo kodiraju kao 0 ili 1. Postupak se rekurzivno nastavlja sve dok se abeceda ne sadrži samo dva znaka, koji se tada jednoznačno kodiraju; prvi s 0, a drugi s 1.

Originalna podjela ulazne abecede koju su predložili Grossi, Grupta i Vitter [W1] abecedu uvijek dijeli na dva jednaka dijela. Podjela abecede koja je implementirana u sklopu ovog projekta razlikuje se od originalne. Naime, kako bi se ujednačila veličina podijeljenih abeceda, a time i memorijsko zauzeće strukture, abecede se dijele na osnovu učestalosti pojavljivanja pojedinog znaka abecede unutar ulaznog niza. U ovoj implementaciji, abeceda se dijeli na prvom znaku čiji broj pojavljivanja u ulaznog nizu, zbrojen s ukupnim brojem pojavljivanja svih znakova koji su u abecedi njegovi prethodnici, veći ili jednak polovici veličine ulaznog niza.

Na slici w1 prikazana je usporedba izgradnje stabla originalnim (gore) i ovdje implementiranim postupkom (dolje). Abeceda ulaznog niza je {a,b,c,d}. Broj pojavljivanja znaka a je 4, znaka b 1, znaka c 5, a znaka d 10 puta. Veličina ulaznog niza iznosi 20. U prvom slučaju, ulazna abeceda dijeli se na podabecede {a,b} i {c,d}, dok se u drugom slučaju abeceda dijeli na {a,b,c} i {d}. Razlog tomu je što je zbroj broja pojavljivanja znakova a, b i c jednak polovici veličine ulaznog niza.

Razlog zašto je ovaj način memorijski efikasniji objašnjen je u idućem poglavlju iz razloga što ovisi o RRR strukturi koja pohranjuje *bit-vektore* stabla.



Slika w1

[W1] R. Grossi, A. Gupta, and J. Vitter. High-order entropy-compressed text indexes. In Proceedings of the 14th annual ACM-SIAM symposium on Discrete algorithms, pages 841–850. Society for Industrial and Applied Mathematics, 2003.

Pronalazak ranga znaka unutar wavelet stabla

Slika w2 prikazuje primjer pronalaska ranga unutar wavelet stabla. Točnije, primjer pronalazi rang znaka c na poziciji 9. U prvom koraku, budući da znamo da je slovo c kodirano s 0, brojimo nule do indeksa 9. Nakon što smo pronašli 7 nula, spuštamo se u slijedeću razinu stabla - u lijevu granu, iz razloga što je znak c kodiran s 0. Sada, u ovoj grani, budući da je znak c kodiran s 1, brojimo koliko ima jedinica u prvih 7 znakova. Važno je primjetiti kako u ovom koraku ustvari idemo do indeksa 6, a ne 7. Nakon što su pronađene 4 jedinice, spuštamo se u iduću razinu, u desnu granu, te brojimo jedinice do četvrtog znaka (indeks 3). U ovoj razini, ujedno i posljednjoj razini stabla, do četvrtog znaka prebrojavamo 3 jedinice iz čega zaključujemo da je rang znaka c na poziciji 9 jednak 3.



Slika w

RRR struktura

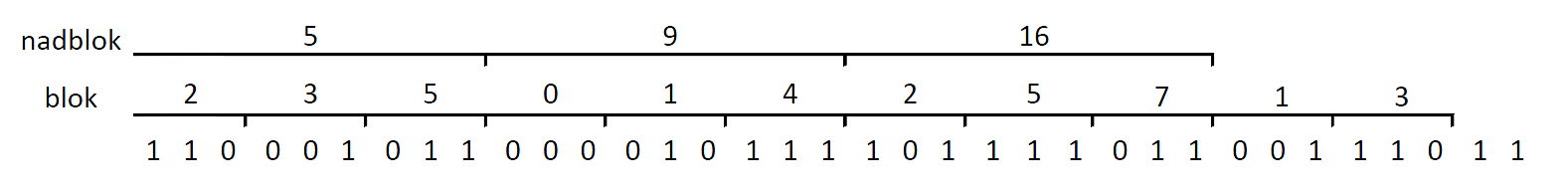
RRR struktura je struktura koja za dani niz nula i jedinica proizvoljne duljine vraća broj postavljenih jedinica do određene pozicije. Općenito, za dobivanje broja jedinica u nizu nula i jedinica, potrebno je proći kroz cijeli niz i provjeravati svaki element niza. Takvo prebrojavanje niza odvija se u linearnoj vremenskoj složenosti O(n) gdje je n duljina niza. RRR strukturom se ukupan broj jedinica do određene pozicije u nizu može izračunati u konstantnom vremenu O(1). Također, korištenjem RRR strukture se postiže i implicitna kompresija.

Ideja RRR strukture je da se inicijalno nad nizom konstruiraju blokovi (engl. *buckets*) i nadblokovi (engl. *superbuckets*) u kojima se pohranjuju brojevi jedinica za određene intervale niza. Na taj način se izbjegava pregledavanje cijelog dijela niza u potrazi za brojem jedinica, već je potrebno pregledati samo mali dio niza i određene vrijednosti pohranjene u blokovima i nadbloku.

RRR struktura implementirana u ovom projektu se djelomično razlikuje od originalne RRR strukture koju su predložili Raman, Raman and Rao [000]. U originalnoj RRR strukturi niz je podijeljen u blokove tako da je svaki blok predstavljen parom (cˇi,oˇi) te su definirane dodatne tri tablice.[001] Implementacija u ovom projektu je nešto drugačije izvedena, ali je zadržala izračun broja jedinica u nizu u konstantnom vremenu te je opisana u nastavku teksta.

Konstruiranje RRR strukture

U ovoj implementaciji RRR strukture, niz duljine n se dijelio u blokove veličine l i nadblokove veličine l^2, a vrijednost l se pritom izračunavala kao logaritam duljine niza, tj. l=logˇ2(n). Stvorena su polja BS i SBS veličina floor(n/l) i floor(n/l^2) za pohranjivanje vrijednosti blokova i nadblokova. Za stvaranje RRR strukture potrebno je proći kroz cijeli niz znamenki. Brojač jedinica je inicijalno postavljen na nula i povećava se kada se prolaskom kroz niz naiđe na element koji ima vrijednost 1. Kada se prođe kroz l elemenata niza, vrijednost brojača pohranjuje se u polje BS kao zbroj vrijednosti stvorenih blokova u još neispunjenom nadbloku. To znači da se blokovi u jednom nadploku pune kumulativno. Kada se prođe l^2 elemenata niza, vrijednost novog nadbloka jednaka je zbroju vrijednosti prošlog nadbloka (ako postoji) i vrijednosti brojača, odnosno zadnjeg konstruiranog bloka. Nakon izračuna vrijednosti nadbloka brojač se postavlja ponovo na vrijednost nula. Primjer ispunjenih vrijednosti u poljima BS i SBS dan je slikom1.

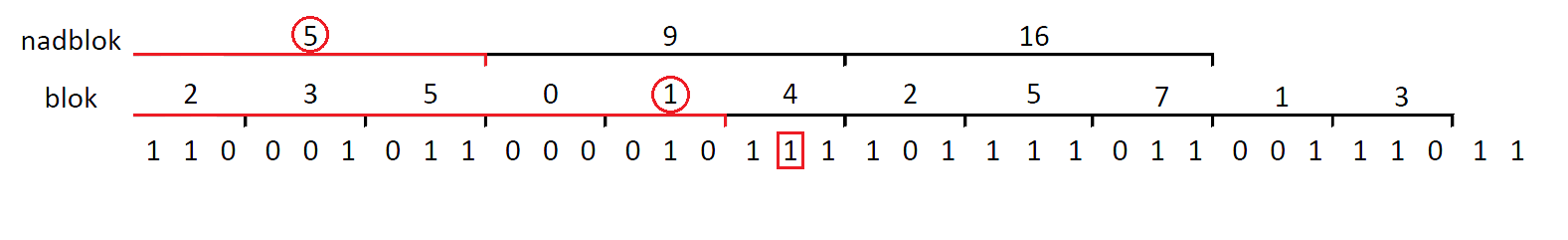


Ukoliko se osvrnemo na primjer wavelet stabla prikazanim u prethodnom poglavlju (slika w1), lako možemo izračunati da, ukoliko stablo izgrađujemo originalnim algoritmom, ukupna veličina izgrađenih polja BS i SBS jednaka je 17. Ako se stablo izgradi dijeljenjem abecede na osnovi broja pojavljivanja znakova unutar ulaznog niza, ukupna veličina izgrađenih BS i SBS polja iznosi 14. Već na ovako malom primjeru zamjećuje se poprilična razlika između dobivenih veličina polja BS i SBS.

Izračunavanje broja jedinica u nizu

Broj jedinica u nizu se izračunava kao zbroj vrijednosti zadnjeg popunjenog nadbloka i bloka do zadane pozicije te broja jedinica u ostatku niza koji nije obuhvaćen blokom. Zbog načina punjenja blokova i nadblokova, kod računanja ukupnog broja pojavljivanja jedinica u nizu do određene pozicije potrebno je pripaziti na određene situacije. Ako se tražena pozicija nalazi na mjestu u nizu jednakom k\*l^2, onda je broj pojavljivanja jedinica u podnizu jednak samo vrijednosti nadbloka, a ako se pozicija nalazi na mjestu u nizu k\*l, to znači da je vrijednost broja jedinica u podnizu dan ili samo nadblokom (ranije naveden slučaj) ili zbrojem vrijednosti nadbloka i bloka, te nema dijela niza koji nije obuhvaćen blokom/nadblokom za koji treba dodatno provjeravati vrijednosti znamenki. U ostalim slučajevima se vrijednostima zadnjeg popunjenog nadbloka i bloka pridodaje broj jedinica u dijelu niza od zadnjeg popunjenog bloka do pozicije do koje se traži izračun broja pojavljivanja jedinica u podnizu, i to je jedini dio u računanju kada je potrebno slijedno prolaziti kroz elemente niza prebrajajući pojavljivanje jedinica.

U nastavku je opisan postupak izračunavanja broja jedinica za primjer niza prikazanog Slikom1. Neka se želi izračunati broj pojavljivanja jedinica do pozicije 17. elementa niza (uključujući). Zadnji popunjeni nadblok moguće je pronaći formulom ind\_nadblok = floor(pozicija/l^2) = floor(17/9) = 1, a zadnji popunjeni blok formulom ind\_blok = floor(pozicija/l) = floor(17/3) = 5. Broj pojavljivanja jedinica u ostatku podniza računa se provjeravanjem broja jedinica u dijelu niza od pozicije ind\_blok\*l+1 do tražene pozicije (uključujući). U ovom primjeru potrebno je provjeriti pojavljivanje jednica u još dva elementa niza, 16. i 17. elementu, čime se dobiva vrijednost dodatnih jedinica dodatne = 2. Ukupan broj pojavljivanja jedninica u nizu do 17. elementa niza se izračunava kao ukupno = SBS(ind\_nadblok) + BS(ind\_blok) + dodatno = 5 + 1 + 2 = 8.



[000] R. Raman, V. Raman, and S. Srinivasa Rao. Succinct indexable dictionaries with applications to encoding k-ary trees and multisets. SODA, str. 233-242, 2002.

[001] Practical Rank/Select Queries over Arbitrary Sequences - Francisco Claude and Gonzalo Navarro (spire08.1-RRR.pdf)

Burrows-Wheeler transformacija

Burrows-Wheeler transformacija(skraćeno BWT), služi za kompresiju tekstualnog niza na osnovi pronalaska ponavljajućih uzoraka unutar tog niza. Ova transformacija je reverzibilna i ne iziskuje pohranu dodatnih podataka za rekonstrukciju originalnog niza. Osim ovoga, prilikom BWT transformacije, ni jedan znak originalnog niza se ne mijenja.

Algoritam transformacije provodi se u 4 koraka:

1. Na kraj teksta postavlja se znak $ koji označava kraj te je, leksikografski gledano, najmanji znak ulazne abecede.
2. Stvara se matrica svih različitih cikličkih pomaka niza dobivenog nakon 1. koraka.
3. Matrica iz koraka 2 se sortira u leksikografskom poretku.
4. Dobivena transformacija isčitava se iz poslijednjeg stupca matrice dobivene u koraku 3.

Tijek algoritma za riječ «banana» prikazan je na slici BWT1.



Slika BWT1

**FM indeks**

Općenito

FM indeks je samostojeći indeks koji su 2000. godine u svom radu opisali i objavili Ferragina i Manzini. Općenito indeks je podatkovna struktura koja omogućuje učinkovito dohvaćanje podataka. Samostojeći indeks (*self-index*) je struktura koja se stvara nad određenim tekstom te ga indeksira na način da je memorijski proporcionalna veličini komprimiranog teksta te omogućava efikasno dohvaćanje i prebrojavanje dijelova tog teksta (bez potrebe za dekompresijom cijelog teksta).

Originalna implementacija

FM indeks count je implementacija FM indeksa koja služi za pronalaženje broja pojavljivanja uzorka P u tekstu S. Općenito se FM indeks temelji na *Burrows-Wheeler transformaciji* i *sufisknom polju*, ali kod izvedbe samo prebrojavanja, ne i dohvaćanja svih pojavljivanja (FM indeks *find*), struktura sufiksnog polja se može izostaviti.

FM indeks count je ustvari samo izvedba algoritma pretraživanja unatrag (engl. *backward search*) u kojemu se kreće s pretraživanjem od zadnjeg znaka uzorka P te ako uzorak postoji u tekstu, algoritam vraća koliko se puta on u tekstu ponavlja.

Pronalaženje broja pojavljivanja uzorka u tekstu provodi se tako da se nad zadanim tekstom prvo stvori FM indeks sa svojim strukturama OCC tablicom i C tablicom te se zatim koristeći stvorene tablice provodi prebrojavanje za zadani uzorak.

U originalnoj implementaciji FM indeksa [002], algoritam pretraživanja unatrag se provodi nad nizom znakova koji je dobiven iz originalnog teksta na koji su primjenjene različite operacije. Nad originalnim tekstom provedene su redom Burrows-Wheeler transformacija, Move-To-Front i Run-length kodiranje te stvaranje prefiksnog koda promjenjive duljine. Za provođenje algoritma potrebne su funkcije C(c) i Occ(c,i). C(c) daje broj znakova u (originalnom) tekstu koji su abecedno prije znaka c uključujući i ponavljanje pojedinih znakova. Occ(c,i) daje broj pojavljivanja znaka c u B[1,i], i=1...|S|, gdje je S originalni tekst, B niz dobiven transformacijama originalnog teksta, a znak c je bilokoji znak iz abecede originalnog teksta. Pokazano je [neka referenca] da brzina izvođenja funkcije Occ(c,i) određuje brzinu izvođenja cijelog algoritma pretraživanja unatrag. U originalnoj implementaciji funkcija Occ(c,i) se računa korištenjem transformiranog (sažetog) originalnog teksta te nekoliko dodatnih pomoćnih struktura. *(opisat te strukture? ako bude ugrubo bit će nejasno, ali sve objasnit će biti predugačko)*

Naša implementacija

U ovoj implementaciji FM indeksa koristi se algoritam pretraživanja unatrag i ranije spomenuta struktura C tablice na isti način kao i u originalnoj implementaciji, ali je funkcija Occ(c,i) izvedena korištenjem Wavelet stabla što će biti detaljnije objašnjeno u nastavku.

Prvi korak je stvaranje C tablice koja za svaki znak iz abecede ulaznog teksta sadrži informacije koliko je znakova abecedno manjih ispred tog znaka abecede, uključujući i ponavljanje znakova. Pretpostavimo da ulazni niz glasi *ananas*. Stvorena struktura (C tablica) izgledala bi ovako:

|  |  |  |  |
| --- | --- | --- | --- |
| c | A | N | S |
| C(c) | 0 | 3 | 5 |

Sljedeći korak je provođenje *Burrows-Wheeler transformacije* nad zadanim tekstom korištenjem *multikey quick* sortiranja. *Burrows-Wheeler transformacija* i *multikey quick* sortiranje opisani su u prethodnim poglavljima. BW transformacijom se iz teksta *ananas* dobiva niz *snnaaa*.

Na temelju dobivenog niza stvara se OCC tablica koja je predstavljena Wavelet stablom. Na temelju izgrađenog stabla, uz korištenje pomoćne RRR strukture, lako se računaju vrijednosti Occ(c,i) koje su potrebne za algoritam pretraživanja unatrag. Struktura Wavelet stabla i RRR struktura su detaljnije opisane u ranijim poglavljima. *DETALJNIJE OPISATI? – vidjeti dal će Maja opisati samo kako se stvara, ili kako se koristi*

Završni korak implementacije je algoritam pretraživanja unatrag koji konačno daje broj pojavljivanja uzorka u zadanom tekstu. Algoritam je implementiran kao i u originalnom radu FM indeksa, prema sljedećem pseudokodu:

*naš pseudokod (Java il C++ sintaksa ?)*

[002] Opportunistic Data Structures with Applications, Paolo Ferragina, Giovanni Manzini, 2000.

Nedostaci implementacije u Javi

Vrijeme, memorija