SVEUČILIŠTE U ZAGREBU

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Bioinformatika

PROJEKT

Improving Bloom Filter Performance on Sequence Data Using k-mer Bloom Filters

Daria Matković

Voditelj: *Mirjana Domazet-Lošo*

Zagreb, sječanj, 2019

Sadržaj

1. Uvod

Bloom filter je učinkovita probabilistička podatkovna struktura koja se koristi za ispitivanje članstva elemenata u skupu. Bloomov filter može dati lažno pozitivan odgovor, tj. neki element može smatrati članom skupa, iako on to nije, ali Bloomov filter sa sigurnošću može provjeriti da neki element ne pripada skupu.

U ovom projektu Bloom filter je služio za spremanje velikog broja k-merova. K-mer je podniz duljine k. K-merovi se nalaze u fasta datoteci koja je preuzeta sa stranice http://bacteria.ensembl.org/index.html.

Za ovaj projekt korištena je već gotova implementacija Bloom filtera koja je dostupna na https://github.com/mavam/libbf/. Gotova implementacija Bloom filtera se automatski instalira.

2. Metode za poboljšanje performansi Bloom Filtra

U ovom projektu su razmatrane dvije vrste poboljšanja performansi Bloom Filtra. Jedan način poboljšanja je tako da se smanjuje broj lažno pozitivnih rezultata, a drugi način poboljšanja performansi Bloom filtera je smanjenjem memorije koju zauzima Bloom filter. Također u oba slučaja želimo da vrijeme potrebno za inicijalizaciju i provjeru članova bude minimalno.

2.1. Metode za smanjenje broja lažno pozitivnih rezultata

Metode za smanjenje broja lažno pozitivnih rezultata su one-sided k-mer Bloom filter i two-sided k-mer Bloom filter. One-sided k-mer Bloom filter osim k-mera koji mu je upit ispituje i njegove susjede, te upitni k-mer smatra članom skupa tek kada je jedan od susjeda (lijevi ili desnii) također član skupa. Two-sided k-mer Bloom filter također provjerava susjede i upitni k-mer smatra članom skupa samo ako su njegovi i lijevi i desni susjedi članovi skupa. Na primjer ako je upitni k-mer: ACCTGATT, onda će lijevi susjed biti XACCTGAT, a desni susjed će biti CCTGATTX, gdje je X={A,C,T,G}. Dakle skup mogućih lijevih susjeda je: AACCTGAT, CACCTGAT, TACCTGAT, GACCTGAT, a skup mogućih desnih susjeda je: CCTGATTA, CCTGATTC, CCTGATTT, CCTGATTG. Za one-sided k-mer Bloom filter potrebno je u skupu pronaći k-mer i jedan od osam susjeda, a za Two-sided Bloom filter potrebno je u skupu pronaći k-mer, jedan lijevi i jedan desni susjed.

2.2. Metode za smanjenje korištene memorije

Za smanjenje korištene memorije potrebno je odabrati samo neke k-merove koji se dodaju u Bloom filter. Postoje 3 metode:

a) Metoda odabirom najboljeg indeksa

Skup spremljenih k-merova se odabire tako da se iz dobivene sekvence sprema svaki s-ti k-mer, počevši od određenog indeksa. Indeks od kojeg se počinju odabirati kmerovi odabran je tako da skup svakog s-tog k-mera koji počinje od tog indeksa treba imati najviše preklapanja sa do sada spremljnim skupom k-merova. Za ovu metodu koristilo se stroga metoda provjeravanja članstva testnog k-mera u skupu spremljenih k-merova.

b) Prorjeđivanje „hitting set” metodom

Koristi se pohlepna metoda kojom se u set dodaju k-merovi koji su se najviše puta pojavili u setu susjeda svih k-merova. Za ovu metodu koristila se opuštena metoda ispitivanja.

c) Metoda odabira svakog s-tog k-kmera iz sekvence

Iz sekvence se svaki s-ti k-mer sprema u Bloom filter. Za ovu metodu koristile su se i opuštena i stroga metoda ispitivanja.

Nakon što spremimo samo određene k-merove u Bloom filter, možemo koristiti dvije metode za provjeru nalazi li se pojedini k-kmer u spremljnom skupu, a to su opuštena metoda provjeravanja i stroga metoda provjeravanja. Na slici 1 prikazani su algoritmi koji opisuju svaku od metoda.

```
1:  function DECIDE_PRESENT(query, Contains_left, Contains_right)
2:      if Contains_right == true and Contains_left == true then
3:          return true
4:      if Contains_right == true or Contains_left == true then
5:          if EDGE_k-mer_SET.contains(query) then
6:              return true
7:      return false

8:  function STRICT-CONTAINS_NEIGHBOURS(query, left_dist, right_dist)
9:      Contains_left ← CONTAINS_SET( S_DISTANT_LEFT_NEIGHBOUR_SET(query, left_dist))
10:     Contains_right ← CONTAINS_SET(S_DISTANT_RIGHT_NEIGHBOUR_SET(query, right_dist))
11:     return DECIDE_PRESENT(query, Contains_left, Contains_right)

12: function RELAXED-CONTAINS_NEIGHBOURS(query, l_dist, r_dist)
13:     Contains_left ← CONTAINS_SET( ⋃_{i≤l_dist} S_DISTANT_LEFT_NEIGHBOUR_SET(query, i))
14:     Contains_right ← CONTAINS_SET( ⋃_{i≤r_dist} S_DISTANT_RIGHT_NEIGHBOUR_SET(query, i))
15:     return DECIDE_PRESENT(query, Contains_left, Contains_right)

16: function STRICT-CONTAINS(query, s)
17:     if BF.CONTAINS(query) then
18:         if STRICT-CONTAINS_NEIGHBOURS(query, s, s) then
19:             return true
20:     for i ← 0 to s − 1 do
21:         if STRICT-CONTAINS_NEIGHBOURS(query, i, s − (i + 1)) then
22:             return true
23:     return false

24: function RELAXED-CONTAINS(query, s)
25:     if BF.CONTAINS(query) then
26:         if RELAXED-CONTAINS_NEIGHBOURS(query, s, s) then
27:             return true
28:     else
29:         for i ← 0 to s − 1 do
30:             if RELAXED-CONTAINS_NEIGHBOURS(query, i, s − (i + 1)) then
31:                 return true
32:     return false
```

Algoritam 1 Algoritam provjere članstva k-mera kod metoda koje prorjeđuju broj spremljenih k-merova

3. Rezultati

Za testiranje je korišteno deset različitih fasta datoteka, različitih duljina. U tablici 1 navedene su imena genoma koja su korištena za testiranje, te broj znakova koji se nalazi u fasta datoteci.

Redni broj	Naziv genoma	Broj znakova u fasta datoteci
1.	Acetobacter ghanensis	17845
2.	Acetobacter pasteurianus	191427
3.	Aster yellows witches'-broom phytoplasma	3972
4.	Candidatus Hepatoplasma crinochetorum	621166
5.	Leptospira borgpetersenii serovar Ballum	361762
6.	Mycoplasma genitalium	580016
7.	Persephonella marina	53682
8.	Rahnella aquatilis	616549
9.	Runella slithyformis	44754
10.	Strawberry lethal yellows phytoplasma	959779

Tablica 1: Fasta datoteke koje su se koristile za testiranje

U nastavku su se zbog jednostavnosti koristili redni brojevi korištenih datoteka, umjesto cijelih naziva. U tablici 2 prikazani su rezultati izvršavanja svih fasta datoteka, sa postavljenom duljinom k-mera 10. Svaka fasta datoteka pokrenuta je 10 puta i uzeta je prosječna vrijednost.

U tablici 2 prikazana je usporedba algoritama s obzirom na vrijeme izvršavanja upita.

Redni broj fasta datoteke	Klasični Bloom Filter	One-sided Bloom Filter	Two-sided Bloom Filter	Metoda odabirom najboljeg indeksa	Hitting set metoda	Metoda odabira svakog s-tog k-mera
1.	0.0002	0.0012	0.0327	0.174	0.174	0.184
2.	0.000214	0.00145	0.3713	1.663	1.401	1.6585
3.	0.000326	0.001386	0.0098	0.0653	0.07	0.0716
4.	0.0002	0.00213	0.9773	4.49	3.52	4.295
5.	0.0002	0.0018	0.6634	2.969	2.349	2.885
6.	0.00018	0.0021	0.814	4.51	3.216	4.269
7.	0.000166	0.00134	0.0992	0.506	0.4543	0.515
8.	0.000213	0.002	1.177	4.939	3.502	4.456
9.	0.0018	0.0013	0.088	0.411	0.38	0.423
10.	0.00017	0.00166	0.8975	5.372	4.297	5.11

Tablica 2: Vrijeme [s] ispitivanja testnog skupa u pojedinoj fasta datoteci

Ispitivanje je provedeno sa 1000 testnih slučajno generiranih k-merova. Može se primjetiti kako je ispitivanje kraće kada su u Bloom filter dodani samo neki k-merovi. Također klasični Bloom filter ima kraće vrijeme ispitivanja od one-sided i two-sided metoda koje uz ispitne k-merove provjeravaju i njihove susjede. U tablici 3 je prikazano kako je također utrošak memorije za spremanje prorjeđenog skupa prosječno manji nego kod spremanja k-merova kod klasičnog Bloom filtera, one-sided i two-sided Bloom filtera.

Redni broj fasta datoteke	Klasični Bloom Filter	One-sided Bloom Filter	Two-sided Bloom Filter	Metoda odabirom najboljeg indeksa	Hitting set metoda	Metoda odabira svakog s-tog k-mera
1.	145.06	145.06	145.06	58.1	138.67	74.04
2.	1195.56	1195.56	1195.56	623.34	152.53	794.47
3.	32.05	32.05	32.05	12.96	30.39	16.46
4.	1870.19	1870.19	1870.19	2027.31	4801.58	2578.08
5.	1963.09	1963.09	1963.09	117.83	2874.29	1501.43
6.	2393.78	2393.78	2393.78	1890.57	4587.54	2407.54
7.	398.34	398.34	398.34	1749.19	422.58	222.79
8.	3405.64	3405.64	3405.64	2007.69	4898.92	2558.92
9.	354.88	354.88	354.88	145.88	353.94	185.74
10.	2413.32	2413.32	2413.32	3130.76	747.66	3983.43

Tablica 3: Memorija [kB] koju zauzimaju spremljeni k-merovi

U tablici 3 vidi se da su metode za prorjeđivanje lošije jer u prosjeku imaju veći broj lažno negativnih rezultata nego druge metode. Najmanji broj lažno negativnih rezultata ima two-sided metoda jer se kod te metode u Bloom filter dodaju svi k-merovi iz fasta datoteke i uz upitni k-mer se provjeravaju i susjedni k-merovi.

Redni broj fasta datoteke	Klasični Bloom Filter	One-sided Bloom Filter	Two-sided Bloom Filter	Metoda odabirom najboljeg indeksa	Hitting set metoda	Metoda odabira svakog s-tog k-mera, relaxed	Metoda odabira svakog s-tog k-mera, strict
1.	22.89	21.15	11.5	8.51	55.94	16.69	15.81
2.	20.13	19.28	12.87	18.57	59.27	26.34	25.89
3.	23.05	20.96	11.27	8.07	53.18	16.76	15.7
4.	13.39	12.99	9.23	21.56	42.25	26.31	26.16
5.	16.73	16.25	11.99	24.49	51.27	26.75	27.59
6.	12.86	12.52	9.44	22.48	42.09	26.98	26.82
7.	20.79	19.42	11.89	13.38	55.77	21.22	20.54
8.	13.59	13.39	10.89	24.88	47.7	30.19	30.06
9.	22.18	20.62	12.1	10.94	58.48	19.23	18.54
10.	12	11.67	8.62	24.39	41.78	28.45	27.4

Tablica 4: Lažno pozitivni primjeri [%]

Rezultati testiranja koji su prikazani u gornjim tablicama mogu se pronaći u mapi „results/testResults” u projektu.

4. Zaključak i usporedba sa originalnom implementacijom

Može se zaključiti kako metode za smanjenje broja lažno pozitivnih rezultata (one-sided i two-sided) po tom kriteriju uspješnije odrađuju ispitivanje od drugih metoda, ali zato je ispitivanje duže. Metode koje se koriste za smanjenje memorije, su u prosjeku po tom kriteriju bolje, ali zato imaju veći broj lažno pozitivnih rezultata.

U originalnoj implementaciji koristile su se drugačije fasta datoteke, drukčija duljina k-merova. U orignalnoj implementaciji je duljina k-merova je bila 20, te je i broj upitnih k-merova bio 1e6. U originalnoj implementaciji korišten je zapis k-merova u obliku znamenaka, dok je u ovom radu korišten originalni zapis iz fasta datoteka, zbog toga je ovaj algoritam lošiji od originalne implemetacije.

5. Literatura

[1] D. Pellow, D. Filippova, C. Kingsford, Improving Bloom Filter Performance on Sequence Data Using k-mer Bloom Filters, Journal of computational biology, Volume 24, Number 6, 2017

[2] Predavanja iz kolegija Napredni modeli i baze podataka:

https://www.fer.unizg.hr/_download/repository/9._NoSQL_4_od_4.pdf

[3] Ulazni podaci (fasta datoteke): http://bacteria.ensembl.org/index.html

[4] Korištena osnovna implementacija Bloom filtera: https://github.com/mavam/libbf/