SVEUČILIŠTE U ZAGREBU

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

SEMINAR

Algoritam za ažuriranje Burrows-Wheelerove transformacije u četiri koraka

Antonio Benc, Matija Herceg, Luka Skukan

Voditelj: *doc. dr. sc. Mirjana Domazet-Lošo*

Zagreb, siječanj, 2016.

Sadržaj

1. Uvod

Burrows-Wheelerova transformacija (BWT) je transformacija teksta, vrlo prikladna za kompresiju. Korištena je u nekim popularnim alatima za kompresiju bez gubitaka, primjerice programu bzip2. Osim pod nazivom Burrows-Wheelerova transformacija, poznata je i pod nazivom *block-sorting compression*.

Konceptualno, tekst nad kojim je izvršena BWT je sličan sufiksnom polju. Zbog te sličnosti BWT se koristi i kao indeksna struktura. BWT teksta T ($bwt(t)$) se često dobiva iz modifikacije sufiksnog polja koja konstrukcija ima $O(n)$ složenost. Pohranjivanje sufiksnog polja u memoriji je jos uvjek glavni problem jer zahtjeva n $n \log n$ bita dok pohrana BWT-a u memoriju zahtjeva $(n \log \sigma)$ bita, gdje je σ broj slova u abecedi.

U ovom seminaru razmatrat će se uobičajne operacije nad tekstom, umetanje znakova, brisanje znakova ili mijenjanje znaka, koje tekst T transformiraju u novi tekst T'. Biti će proučavan utjecaj tih operacija na $bwt(T)$ i biti će predložen algoritam za pretvorbu $bwt(T)$ u $bwt(T')$.

2. Burrows-Wheelerova transformacija

Neka je tekst $T = T[0..n]$ riječ duljine $n + 1$ s abecedom Σ, pri čemu je abeceda Σ konacne velicine σ. Zadnji znak u rijeci T je jedinstveni znak $\$$ (*sentinel*) koji ima vrijedmost manju od svih ostalih znakova u abecedi. Podniz koji pocinje na poziciji i i zavrsava na poziciji j je oznacen s $T[i..j]$, znak na poziciji i je oznacen s $T[i]$ te je ciklički pomak reda i, $T[i..n]T[0..i-1]$, oznacen s $T^{[i]}$.

Burrows-Wheelerova transformacija od T, oznacena s $bwt(T)$, je tekst duljine n+1 koji odgovara zadnjem stupcu, (L), matrice čiji su reci leksikografski sortirani ciklički pomaci $T^{[i]}$. Prvi stupac matrice, (F), je sortiran, tako da se jednostavno može izračunati iz stupca L. Redovi sortiranih cikličkih pomaka, π jednaki su sufiksnom polju od T. Iz toga slijedi kako su sufiksno polje $SA[i]$ i L povezani jednostavnom formulom $L[i] = T[(SA[i] - 1) \mod |T|]$.

$T^{[0]}$	A	T	G	C	G	$
$T^{[1]}$	T	G	C	G	$	A
$T^{[2]}$	G	C	G	$	A	T
$T^{[3]}$	C	G	$	A	T	G
$T^{[4]}$	G	$	A	T	G	C
$T^{[5]}$	$	A	T	G	C	G

$\Rightarrow$

$T^{[5]}$	$	A	T	G	C	G
$T^{[0]}$	A	T	G	C	G	$
$T^{[3]}$	C	G	$	A	T	G
$T^{[4]}$	G	$	A	T	G	C
$T^{[2]}$	G	C	G	$	A	T
$T^{[1]}$	T	G	C	G	$	A
	F					L

$\Rightarrow$

i	π
0	5
1	0
2	3
3	4
4	2
5	1

Tablica 1.: Burrows-Wheelerova transformacija niza $ATGCG\$$

Tablicom 1. prikazana je Burrows-Wheelerova transformacija niza znakova $ATGCG\$$. U tablici lijevo prikazani su svi ciklički pomaci tog niza. U tablici na sredini ti ciklički pomaci su leksikografski sortirani i oznaceni su stupci F i L. U tablici desno prikazan je niz brojeva koji predstavlja redove sortiranih cikličkih pomaka, ujedno i sufiksno polje od niza.

Burrows-Wheelerova transformacija sadrži samo zadnji stupac sortirane matrice cikličkih pomaka L. Za rekonstrukciju počenog niza niza T iz $bwt(t)$ koristi se veza između stupca L i stupca F. Ako znakovima u primjeru s $ATGCG\$$ svakom znaku dodamo broj koji oznacava redni broj pojavljivanja tog slova dobivamo $A_1T_1G_1C_1G_2\$_1$. Tablicom **??** prikazana je BWT niza s rangiranim znakovima zajedno sa stupcima L i

$T^{[5]}$	$\$_1$	A_1	T_1	G_1	C_1	G_2
$T^{[0]}$	A_1	T_1	G_1	C_1	G_2	$\$_0$
$T^{[3]}$	C_1	G_2	$\$_1$	A_1	T_1	G_1
$T^{[4]}$	G_2	$\$_1$	A_1	T_1	G_1	C_1
$T^{[2]}$	G_1	C_1	G_2	$\$_1$	A_1	T_1
$T^{[1]}$	T_1	G_1	C_1	G_2	$\$_1$	A_1

Tablica 2.: Burrows-Wheelerova transformacija niza $ATGCG\$$ s rangiranim znakovima

F. Rotacije koje počinju istim slovom, u primjeru slovom G, sortirane su po znakovima iza tog slova. Kada se te rotiacij ciklički rotiraju za jedno mjesto, slova G će se naći u zadnjem stupcu, dok druga slova biti na početku niza i određivati će leksikografski poredak tih rotacija. Upravo zato je redosljed istih slova u prvom stupcu jednak redosljedu istih slova u zadnjem stupcu. Ovo svojstvo očuvanosti poretka istih znakova u prvom i zadnjem stupcu BWT-a naziva se LF-mapiranje (last to first mapping) i omogućuje rekonstrukciju pocetnog niza iz BWT-a. Formalnije, LF-mapiranje opisuje vezu, odnosno mapiranje, zadnjeg i prvog stupca u listi leksikografski sortiranih cikličkih rotacija niza S, a temelji se na sljedećem: i-ta pojava znaka c u zadnjem stupcu (stupcu L) leksikografski sortiranih cikličkih rotacija odgovara i-toj pojavi znaka c u prvom stupcu (stupcu F). LF mapiranjem, iz BWT-a i uz prisutnost stupca F, moguće je izgraditi početni niz. Stupac F može se dobiti tako da se leksikografski poredaju svi znakovi u stupcu L, odnosno BWT-u. Formula za LF mapiranje znaka na poziciji p glasi

$$LF(p) = C_T[L[p]] + rank_{L[p]}(L, p) - 1, \tag{1}$$

gdje je C_T broj znakova u nizu manjih od znaka $L[p]$, $rank_{L[p]}(L, p)$ broj pojavljivanja znaka $L[p]$ u L do pozicije p.

Stupac L, odnosno BWT je konceptualno blizak sufiksnom polju, a povezuje ih jednostavna transformacija. Stoga se većina algoritama za konstrukciju BWT-a bazira na postoječim algoritmima za računanje sufiksnog polja s kompleksnošću $O(n)$ i primjeni transformacije sufiksnog polja u BWT.

Jednostavna transformacija niza T u niz T' uzrokuje da se BWT od T' mora računati od nule. U nastavku seminara biti će proučeno kako jednostavne operacije nad nizom T utječu na $bwt(T)$.

3. Zaključak

4. Sažetak