XII Telekomunikacioni forum TELFOR 2004, Beograd, Sava Centar, 23.-25.11.2004.g.
Rad za studentsku sekciju

PROBABILISTIČKA ANALIZA KRATKIH SEKVENCI

Slađana Tošić
Fakultet tehničkih nauka u Novom Sadu

I UVOD

Bilo koji prenosni sistem koji koristi vremenski multipleks, zahteva postojanje markera koji pokazuje gde je početak, što omogućava ispravno demultipleksiranje. Za to se obično koristi sekvenca koja se naziva sinhronišuća sekvenca, marker ili sinhro-reč. Optimizaciji markera se ne posvećuje suviše pažnje [1] jer i loš marker može ukazivati na početak, jedino što su performanse sistema malo lošije. Umesto markera može se koristiti inherentna redundansa samog signala. To je, međutim, manje pouzdan način i daje slabije rezultate.

Ovaj rad se bavi problemom pronalaženja izabrane sekvence u nizu podataka koji su slučajni, ali nemaju jednaku verovatnoću pojavljivanja. Sekvence koje se posmatraju su kratke, četiri do osam bita.

Rad obuhvata četiri dela:

-prvi deo se odnosi na raspodelu vremena pretrage T;

-drugi deo se bavi određivanjem verovatnoće nailaska na izabranu sekvencu u okviru slučajnog niza podataka;

-u trećem delu tema je određivanje verovatnoće nailaska na neku od zadatih sekvenci u nizu slučajnih podataka, a da prethodno nismo naišli ni na koju drugu traženu sekvencu;

-četvrti deo govori o određivanju sume verovatnoća simulacije P_{TS}.

II RASPODELA VREMENA PRETRAGE

Verovatnoća simulacije sinhro-reči na određenoj poziciji u ramu zavisi od ishoda prethodnih testova za sve sinhro-reči duže od jednog bita. Za određivanje verovatnoće neophodno je poznavanje raspodele vremena pretrage za sekvencom određene dužine i strukture.

Rad P.T.Nielsena [2], koji govori o očekivanom vremenu pretrage izabrane sekvence u nizu slučajnih podataka, nezaobilazan je pri svakoj analizi ovog procesa. Ovo je inicijalni rad koji je uveo analitičku metodologiju baziranu na bifiksima, ali je ograničen na slučaj podataka iste verovatnoće.

Posmatra se potraga za određenim paternom u nizu slučajnih podataka neograničene dužine.

Neka je A_L alfabet od L simbola, gde je L≥2. Sekvenca koju posmatramo je $\boldsymbol{d}=[d1,d2,d3...]$, gde su svi simboli nezavisni sa verovatnoćom pojavljivanja 1/L. Niz $\boldsymbol{d}$ se naziva *niz slučajnih podataka*. Poslednji elementi niza su :

$\boldsymbol{d}[i,j) = [d_i,d_{i+1}, \ldots ,d_{j-1}]$, za svako $i \geq 0$ i $j \geq i+1$.

Neka je $\boldsymbol{p} = [p_1,p_2, \ldots ,p_n]$ patern od $\boldsymbol{n}$ simbola izabranih iz alfabeta A_L.

Definiše se slučajna promenljiva x takva da je $x=i$ ako i samo ako je :

$\boldsymbol{d}[i, i+n) = p$

$\boldsymbol{d}[j, j+n) \neq p, \qquad 1 \leq j < i$

x je broj pozicija koje su prošle pre nego što se naišlo na patern.

Definicija: Sekvenca $\boldsymbol{a} = [a_1,a_2, \ldots ,a_m]$ $(1 \leq m < n)$ je bifiks od $\boldsymbol{p} = [p_1, p_2, \ldots ,p_n]$ ako je $[p_1, p_2, \ldots , p_m] = [p_{n-m+1}, p_{n-m+2}, \ldots , p_n] = \boldsymbol{a}$

Svih m simbola iz $\boldsymbol{a}$ su istovremeno i prefiks i sufiks od $\boldsymbol{p}$.

Definišu se *bifiks indikatori* h_i od $\boldsymbol{p}$, $1 \leq i < n$ takvi da je $h_i = 1$ ako je $[p_1 ,p_2 , \ldots ,p_i]$ bifiks od $\boldsymbol{p}$, a u suprotnom je $h_i = 0$. Ako je $h_i = 0$ za svako $1 \leq i < n$, kažemo da je $\boldsymbol{p}$ sekvenca *bifix-free, tj.* ona je bez bifiksa, izuzev onih koji su po definiciji jednaki jedinici, a to su $h_0 = h_i = 1$.

Pojam *bifiks* označava podsekvencu koja je istovremeno i prefiks i sufiks. Na primer, binarna sekvenca 10110110 dužine $n = 8$ ima dva bifiksa: 10 i 10110. Bifiks indikator h_i ravan je 1 ako u izabranoj sekvenci (*paternu*) postoji bifiks dužine i, pa je za posmatrani primer $h_0 = h_2 = h_5 = h_8 = 1$ jer je prema konvenciji $h_0 = h_n = 1$. Vrednost ostalih bifiks indikatora za posmatrani primer ravna je nuli.

Teorema : Za bilo koji patern $\boldsymbol{p}$ dužine n ($n \geq 1$), čiji su bifix indikatori $h_1, h_2, \ldots, h_{n-1}$,u nizu slučajnih L-arnih podataka, očekivano vreme pretrage za izabranom sekvencom $\boldsymbol{p}$, ili matematičko očekivanje, iznosi:

$$E\{x\}=\sum_{i=0}^{n} h_i \cdot L^i - n \qquad (1)$$

gde je po definiciji $h_0 = h_n = 1$.

Osim ove teoreme, Nielsen je izveo i zaključak, koji važi za svaku sekvencu $\boldsymbol{p}$ dužine n, a glasi:

$$L^n - n + 1 \leq E\{x\} \leq \frac{L^{n+1}-1}{L-1} - n \qquad (2)$$

Jednakost na levoj strani važi ako i samo ako je $\boldsymbol{p}$ sekvenca bez bifiksa, izuzev onih po definiciji, a jednakost na desnoj strani važi ako i samo ako se sekvenca $\boldsymbol{p}$ sastoji od n istih simbola iz alfabeta A_L.

Njegovo zapažanje je da prisustvo bifiksa u paternu utiče na njegovo kašnjenje u nizu slučajnih podataka.

Prethodne formule važe za slučaj jednakih verovatnoća pojavljivanja simbola ***p***. U ovom radu se razmatra verovatnoća nailaska na jednu fiksiranu sekvencu u nizu podataka koji nisu iste verovatnoće. Ako je reč o binarnim podacima, verovatnoća pojave jedinice je p, a verovatnoća pojave nule je q. Ako posmatramo opšti slučaj, gde su u pitanju L-arni podaci, verovatnoće pojedinih simbola (na primer simbola 0, 1, 2, . . . , L-1) su $p_0, p_1, p_2, ..., p_{L-1}$.

Odgovarajuće matematičko očekivanje (vreme pretrage) je:

$$T = 1 - N \cdot r_N + \sum_{m=0}^{N-1} h_{N-m} \cdot r_m \qquad (3)$$

N predstavlja dužinu sekvence, h_i su bifiks indikatori, koji su već objašnjeni u prethodnom izlaganju, a novina je parametar r_m, koji predstavlja verovatnoću ostatka sekvence, ako bifiks dužine N-m postoji.

To se može objasniti na primeru:

SEKVENCA: 0010100, N=7

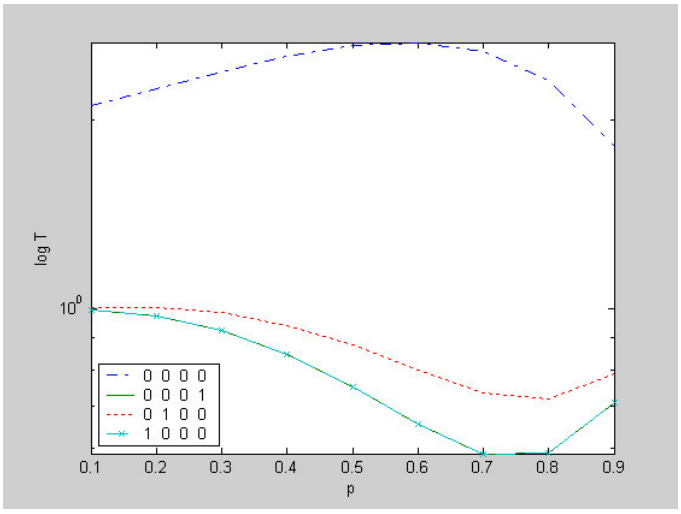
Verovatnoća sekvence koja ima 2 jedinice i 5 nula je $\Pr\{1\} = p^2q^5$

Ima dva bifiksa – 0 i 00.

h_1=1 – ostatak sekvence je 010100 i ta verovatnoća je $r_6 = p^2q^4$ i h_2=1 - ostatak sekvence je 10100 i ta verovatnoća je $r_5 = p^2q^3$.

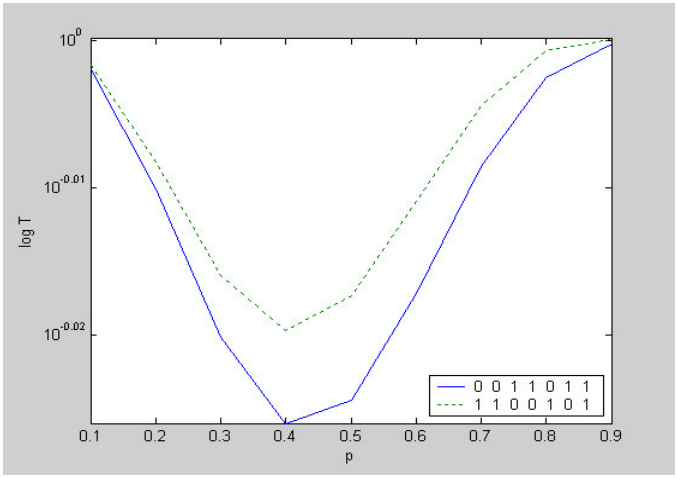
Znamo da je h_0=1 po default-u, odgovarajuce $r_N = r_7 = p^2q^5$.

Znamo da je h_N=1, odgovarajuce po defaultu r_0=1.



Slika 1. Raspodela vremena pretrage za neke od četvorobitnih sekvenci

Brojni grafici, koji su dobijeni u ovom istraživanju, potvrđuju Nielsenov zaključak, koji važi i za sekvence kod kojih verovatnoća pojave pojedinih simbola u nizu nije ista. Najveće očekivanje T dostiže se za sekvence koje u svom sastavu imaju iste simbole (sve nule ili sve jedinice). Što je manji broj bifiksa, i vreme pretrage dostiže manje vrednosti, tako da je najmanje kod sekvenci bez bifiksa. To se može videti na slici 1, gde su prikazane raspodele vremena pretrage za neke četvorobitne sekvence. Sekvenca sa svim nulama dostiže najveće vrednosti, sekvenca 0100, koja ima jedan bifiks, uzima manje vrednosti za T, a sekvence 0001 i 1000, koje su bez bifiksa, dostižu najmanje vrednosti. Njihovi grafici se poklapaju zbog rasporeda nula i jedinica u njima. Slika 2 takođe pokazuje da je i u slučaju sedmobitnih sekvenci vreme pretrage manje što je broj bifiksa manji. Sekvenca 0011011 je bez, a sekvenca 1100101 ima jedan bifiks.



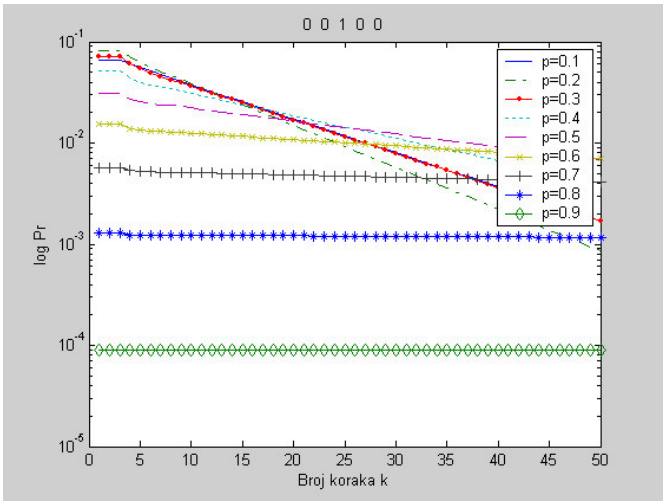
Slika 2. Raspodela vremena pretrage za dve sedmobitne sekvence

III VEROVATNOĆA NAILASKA NA ZADATU SEKVENCU

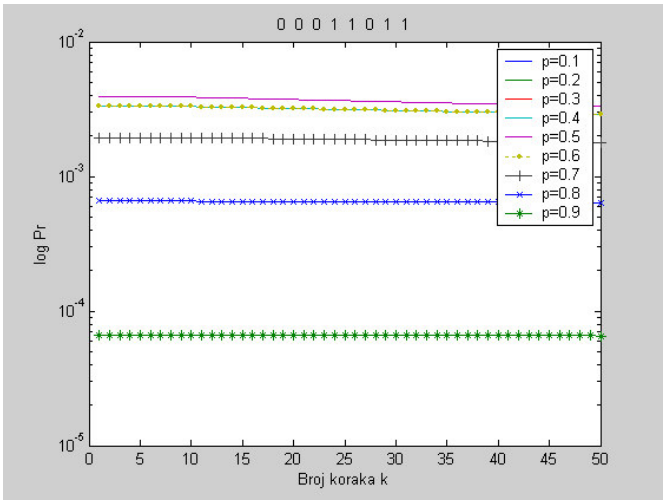
Tema ovog dela rada je određivanje verovatnoće nailaska na zadatu sekvencu, koja je i ovde kratke dužine, u neograničenom nizu podataka. Formula, koja će ovde biti navedena, nadovezuje se na prethodna izlaganja, tako da je ovde prikazana bez dodatnih objašnjenja, koja su data u prethodnom delu. Formule (3) i (4) još nigde nisu objavljene.

$$\Pr\{k\} = \sum_{m=1}^{\min(k-1,N)} \left(h_{N+1-m} \cdot r_{m-1} - h_{N-m} \cdot r_m\right) \cdot \Pr\{k-m\} \qquad (4)$$

Pojam bifiksa h, kao i pojam repa sekvence r, već su objašnjeni, pa neće biti ponovo navođeno njihovo značenje. Parametar k predstavlja broj koraka pretrage. Posmatra se u opštem slučaju L-arni alfabet, tako da su verovatnoće pojave simbola (na primer simbola 0, 1, 2, ..., L-1) $p_0, p_1, p_2, ..., p_{L-1}$. Kada se radi o binarnim podacima, verovatnoća pojave jedinice je p, a verovatnoća pojave nule je q. Ono što ranije nije bilo naglašeno, a što se podrazumeva da važi za sve proračune ovde izvršene kada je u pitanju binarni alfabet, je da je zbir ovih verovatnoća jednak jedinici, tj. p+q=1.



Slika 3. Raspodela verovatnoća Pr za petobitnu sekvencu pri različitim verovatnoćama pojave jedinice *p*, odnosno pojave nule *q*



Slika 4. Raspodela verovatnoće Pr za osmobitnu sekvencu pri različitim verovatnoćama pojave jedinice *p*, odnosno pojave nule *q*

Slike 3 i 4 prikazuju raspodelu verovatnoće Pr u prvih pedeset koraka pretrage kada verovatnoće pojave jedinice, odnosno nule uzimaju neke od vrednosti iz skupa verovatnoća od 0.1 do 0.9. Na slici 4 pojedine krive se poklapaju zato što sekvenca sadrži jednak broj jedinica i nula (po četiri), pa će se za određene verdnosti *p* i *q* grafici poklapati. Primećeno je da kod sekvenci iste dužine, sa jednakim brojem nula i jedinica u njima, neće uvek doći do poklapanja grafika za odgovarajuće vrednosti za *p* i *q* zbog prisustva različitog broja bifiksa u sekvencama. Što su sekvence duže, dobijaju se ravnomernije rapodele za Pr tokom svih koraka pretrage, odnosno krive imaju manje nagibe.

IV PROCES PRETRAGE ZA GRUPU SEKVENCI

Tema ovog dela rada je proces pretrage za grupu fiksnih sekvenci u nizu slučajnih podataka neograničene dužine, pri čemu se pretraga zaustavlja pri nailasku na bilo koju od zadatih sekvenci, a da prethodno nije pronađena ni jedna druga tražena sekvenca.

Cilj višestruke pretrage je pronalaženje bilo koje od *M* zadatih sekvenci, koje se kreću duž niza slučajnih podataka. Svaka sekvenca se sastoji od *N* *L*-arnih simbola.

Proces pretrage počinje (prvi test za *k*=1) poređenjem prvih *N* primljenih simbola sa svakom od *M* sekvenci. Test je neuspešan ako se ni jedna od *M* sekvenci ne poklapa sa primljenih *M* simbola. U tom slučaju se mesto pozicioniranja pomera za jedan, i sada se porede simboli od drugog do *N*+1 sa svakom od *M* sekvenci. Pretraga se završava ako je k-ti test uspešan, što znači da je primljeni niz podataka dužine k+N-1 zadovoljio *širi uslov poklapanja*: poslednih *N* simbola niza podataka jednako je jednoj od *M* sekvenci, pri čemu ni jedna od ovih *M* sekvenci nije pronađena ni na jednoj poziciji ranije. Verovatnoća ovog događaja (verovatnoća da je broj testova tačno *k*, ili verovatnoća da je jedna od *M* sekvenci pronađena pri *k*-tom testu) je:

$$\Pr\{k\} = \sum_{i=1}^{M} \Pr^{(i)}\{k\} = \sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{m=1}^{\min(k-1,N)} (h_{ji}^{(N+1-m)} \cdot r_i^{(m-1)} - h_{ji}^{(N-m)} \cdot r_i^{(m)}) \cdot \Pr^{(j)}\{k-m\} \tag{5}$$

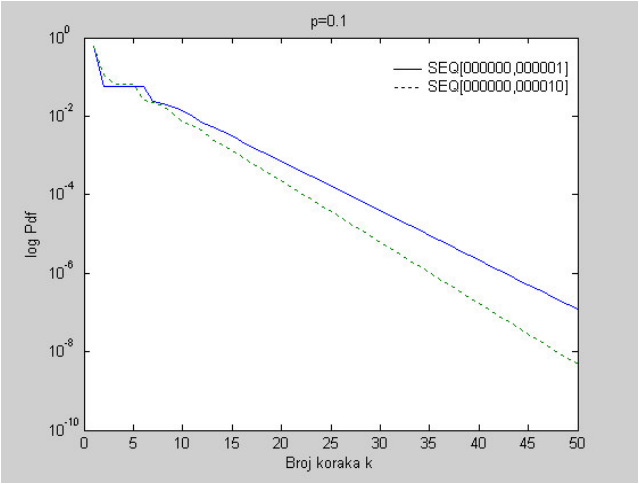
Parametar $h_{ij}^{(n)}$ zove se kros-bifiks indikator, a njegova definicija je: $h_{ij}^{(n)} = 1$ ako je *n* poslednjih bita sekvence br. *i* jednako *n* prvih bita sekvence *j*. Na primer, binarne (*L*=2) sekvence od kojih je *i*-ta ravna 0**001** a *j*-ta ravna **001**1 imaju 3-bitni cross-bifix $h_{ij}^{(3)} = 1$, a očigledno obrnuto ne važi, tj. $h_{ji}^{(3)} = 0$. Po default-u:

$$h_{ij}^{(0)} = 1, \quad h_{ij}^{(N)} = \begin{cases} 0, i \neq j \\ 1, i = j \end{cases}, \quad i, j = 1, \Lambda, M \cdot \tag{6}$$

Na slici 5 prikazane su raspodele ove verovatnoće pri potrazi za po dvema šestobitnim sekvencama, pri čemu je uzeto da je verovatnoća nailaska na jedinicu p=0.1, a analogno tome q=0.9. U jednom slučaju tragamo za sekvencama 000000 i 000001, a u drugom za 000000 i 000010. Grafici se ne poklapaju zbog prisustva različitog broja bifiksa, sekvenca 000001 nema, a sekvenca 000010 ima jedan bifiks, mada obe imaju po jednu jedinicu i pet nula.

V SUMA VEROVATNOĆA SIMULACIJE P_{TS}

U okviru niza podataka posmatramo fiksni ram i podatke unutar njega. Upoređujemo sekvencu koja nailazi sa sekvencom u ramu. Ukoliko se simboli, koji se upoređuju, poklapaju, dodeljujemo im verovatnoću Qe, a ako se razlikuju, dodeljujemo verovatnoću Pe, koja se naziva *verovatnoća greške.*



Slika 5. Raspodela verovatnoća nailaska na jednu od dveju šestobitnih sekvenci

Mora važiti uslov:

Qe = 1-Pe .

Simboli, koji se porede sa simbolima van rama, uzimaju vrednost *p* ako su jedinica, ili *q* ako su nula, bez obzira na simbole koji se nalaze u nizu podataka u kom je prisutan ram.

Posmatrajmo primer jednog niza podataka, u kome je fiksiran ram, i jedne sekvence(npr. 0110).

x	x	x	0	1	1	0

0	1	1	0

Slika 6. Računanje prvog sabirka za Pts

U prvom koraku samo se jedan bit pristižuće sekvence poredi sa samo jednim bitom u ramu. U ovom slučaju oba bita su nule, tako da će prvi član sume glasiti: $p^2 \cdot q \cdot Q_e$ .Proces poređenja se dalje nastavlja tako što se porede po dva bita iz sekvence i iz rama, zatim po tri, a krajnji rezultat je:

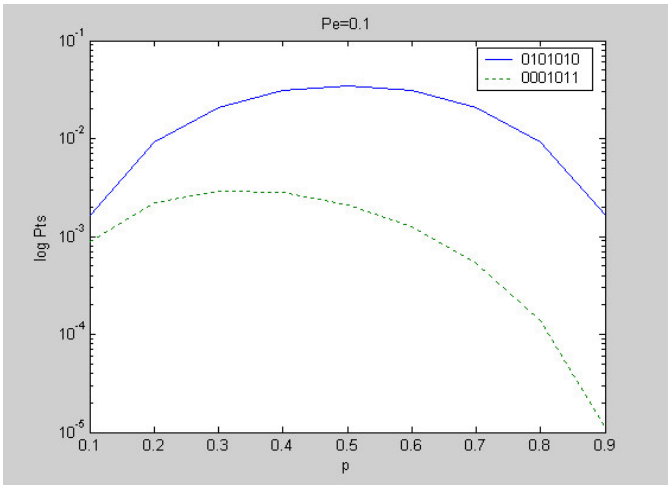
$$P_{TS} = p^2 \cdot q \cdot Q_e + q \cdot p \cdot P_e^2 + Q_e \cdot P_e^2 \cdot q \qquad (7)$$

P_{Ts} je veličina koja može biti i veća od jedinice.

I ova veličina je zavisna od prisustva bifiksa u sekvencama. Sekvence bez bifiksa dostižu najniže vrednosti, dok veći broj bifiksa znači veće vrednosti verovatnoće, što predstavlja analogiju ove veličine sa vremenom pretrage T. Kod sekvenci bez bifiksa ova verovatnoća raste pri porastu verovatnoće greške, kao i pri porastu dužine sekvence, dok kod sekvenci sa bifiksima lagano opada sa porastom verovatnoće greške.

VI ZAKLJUČAK

Ovaj rad je proistekao iz diplomskog rada, koji je urađen pri Katedri za telekomunikacije i obradu signala na Fakultetu



Slika 7. Raspodela sume verovatnoća simulacije P_{TS} za neke od sedmobitnih sekvenci

tehničkih nauka u Novom Sadu. Zadatak rada je bio da se ispita ponašanje statističkih osobina kratkih sekvenci u slučaju kada verovatnoća pojedinih simbola nije podjednaka, za šta do sada nije postojao matematički aparat.

Zaključili smo da statistički parametri,koji opisuju sekvencu u procesu pretrage,zavise od verovatnoće,ako i od prisustva bifiksa u sekvencama.Prikazani primeri su na binarnom nivou (L=2),ali se očekuje da će istraživanja nad višenivoskim sekvencama dati još interesantnije rezultate.

LITERATURA

[1] H. Huh, T. Pande, J. Krogmeier: "*Decoder'Assisted Frame Szynchroniyation in the Presence of Phase – frequency Noise*", submitted to ICC2005, South Korea

[2] P.T. Nielsen: "*On the Expected Duration of a Search for a Fixed Pattern in Random Data*", IEEE Trans. on Inf. Theory., pp. 702-704, Vol. IT-19, September 1973.

[3] D. Bajić, D. Drajić: "*Duration of search for a fixed pattern in random data: Distribution function and variance*", Electronics Letters, 1995, Vol. 31. No. 8, pp 631-632.

[4] D. Bajić, J. Stojanović: "*Distibuted Sequences and Search Process*", IEEE Internation Conference on Communication ICC2004, Paris, France, June 2004, CT08-6.

[5] D. Bajić, J. Stojanovic and J. Lindner: "*Multiple Window-sliding Search*", IEEE International Symposium on Information Theory (ISIT-2003), Yokohama, Japan, June 2003, pp. 249.

Zahvalnica: Ovaj rad je završen zahvaljujući velikoj pomoći prof. dr. D. Bajić.

Abstract: This paper presents results of given formulas considering statistical parameters of search process. These results were gathered by simulation and presented in graphical form. For these analyses was used MATLAB programing language.

PROBABILITY ANALISE OF SHORT SEQUENCES, Tošić Slađana.