Šifra zasnovana na statističkim osobinama teksta

Luka Milinković

***Sadržaj* — U radu je prikazana jedna ideja sistema za šifrovanje podataka, zasnovanog na statističkim osobinama teksta. Prvo je opisan algoritam i programsko rešenje, a zatim je preko primera prikazana realizacija šifre. Na kraju rada je analiziran kriptogram dobijen šifrovanjem teksta i na njemu je pokazano koliko su, ustvari, sami podaci iz poruke dobro skriveni.**

***Ključne reči* — Zaštita, kriptogram, poruka, tajnost, šifra.**

I UVOD

Potreba ljudi za čuvanjem i prenošenjem podataka, bilo da su to lični, tekstualni, kao što su pisma, fotografije ili video, ili poverljivi sadržaji, vojni ili civilni, Vladini ili ne, doprineli su razvoju tehnika za njihovu tajnost. Pojavom novih i naprednih tehnologija, računara i Interneta, tajnost je dovedena u pitanje. Sada svako ko i malo zna o računarima lako može da dođe do tih podataka i da ih iskoristi. Zadatak sigurnosnih sistema je borba protiv neovlašćenih pristupa tajnim sadržajima. Zato se konstruišu kriptosistemi koji će posle šifrovanja poruke stvoriti kriptogram, koji će biti nečitljiv za one osobe koje ne treba da imaju pristup tim podacima [1]. Šifrovana poruka ne sme biti nerazumljiva samo na prvi pogled, već treba da ostane tajna i posle kriptoanalize. Pokušaj razbijanja, različitim metodama, nikako ne treba da olakša posao kriptoanalitičarima, ali bi bilo dobro da im oteža.

Šifra koja je predložena u ovom radu baš u tome uspeva. Ona u potpunosti narušava statističku zavisnost među simbolima u tekstu i na taj način „mrsi račune“ kriptoanalitičarima. Svaki put kada pomisle da su došli do rešenja vratiće se ustvari na početak ili će biti navedeni na pogrešan zaključak.

II STATISTIČKE OSOBINE TEKSTA

Svaki tekst se sastoji od manjeg ili većeg broja karaktera. Oni mogu biti: slova, brojevi, razmak, interpunkcijski znaci ili neki drugi znaci iz ASCII (*American Standard Code for Information Interchange*) tabele [2]. Neki od njih se u tekstu moraju pojavljivati više puta, kao što su to, na primer samoglasnici i razmak, a neki manje, kao što je slučaj sa, na primer, slovima đ i dž i znakom pitanja.

Postoji još jedna bitna osobina teksta. Sama struktura bilo kog pisma, odnosno jezika, je takva da reči koje se koriste u njemu nemaju raspored slova na bilo koji način, već na tačno određen, njemu karakterističan. Na primer, naš jezik ne poznaje reči u kojima su dva slova A ili slova Đ i Dž jedno do drugog. Ovo znači da u svakom pismu postoji statistička zavisnost između slova. Pored toga i struktura svakog teksta je takva da ne može posle bilo kog karaktera da dođe bilo koji karakter već samo neki, kojih je moglo bi se reći malo (najviše 30). Ovakav je slučaj i sa, na primer, interpunkcijskim znakom tačka. Posle nje mogu doći samo razmak ili novi red. Ova osobina teksta unosi dodatnu statističku zavisnost među karakterima.

Baš na ovim, do sada opisanim, osobinama teksta zasniva se algoritam objašnjen u nastavku.

III MOJA ŠIFRA

A. Opis algoritma

Priložena šifra je zamišljena kao metod zaštite tekstualnih poruka i njihovom bezbednom čuvanju ili prenošenju kroz nesigurne kanale.

Algoritam se sastoji iz dva dela. Prvi deo je obavezan i može se podeliti u tri faze. U prvoj fazi se određuje koliko se puta koji karakter pojavljuje u otvorenom tekstu i na osnovu toga se izračunava njegova verovatnoća pojavljivanja. U drugoj fazi se formira grupa različitih simbola, odnosno kodova, koji će da zamene ove karaktere. Za kodove se uzimaju decimalni brojevi. Sada se svakom karakteru dodeljuje određeni broj simbola, koji zavisi od verovatnoće pojavljivanja karaktera u tekstu. Na primer, ako je verovatnoća pojavljivanja nekog simbola 5% onda će se od 100 različitih simbola njemu dodeliti 5. Svaki simbol se može dodeliti samo jednom, čime se formira baza kodova. Kad god se u tekstu pojavi neki karakter, u kriptogramu će se pojaviti neki od njegovih kodova. Na ovaj način se formira šifrovani tekst u kome se svaki simbol pojavljuje skoro isti broj puta. Do sada je bilo lakše raditi sa decimalnim brojevima, ali u trećoj fazi se oni prebacuju u binarni oblik zbog lakšeg čuvanja, odnosno slanja kroz „mrežu“. Sve binarne reči su iste dužine, da bi se znalo gde koja počinje, a gde se završava, i povezane su u jedan niz. U ovom binarnom nizu ne postoji statistička zavisnost, a nule i jedinice se pojavljuju približno isti broj puta.

U drugom delu, koji je opcioni i doprinosi usavršavanju koda, niz bita se deli na delove jednakih dužina, koji se nazivaju kodne reči. Tako dobijene kodne reči ne smeju biti dužine veće od 8 bita, jer je 256 najveći mogući broj različitih karaktera iz ASCII tabele. Svaka kodna reč se zamenjuje sa njoj odgovarajućim karakterom ili sa onim iz

Luka B. Milinković, dipl. inž., Elektrotehnički fakultet iz Beograda; elektronska adresa: lukaui@yahoo.com

ASCII tabele koji se odredi. Ne moraju se iskoristiti svi karakteri iz ove tabele, ali se vodi računa da broj različitih karaktera bude stepen dvojke, 256, 128, 64, 32 ili 16, jer je stepen ustvari dužina kodne reči. Ovako se dobija kriptogram u kome je verovatnoća pojavljivanja simbola u velikoj meri ujednačena. Statistička zavisnost, takođe, ne postoji, jer se sada karakteri mogu pojaviti na sasvim slučajan način.

B. Programsko rešenje

Postoje tri stvari koje je važno dobro realizovati u programu kako bi se formirao sistem kao što je opisano u prethodnoj celini:

- odabrati kodove,
- dodeliti ih karakterima i formirati bazu kodova i
- šifrovati tekst, odnosno zameniti karaktere kodovima.

Prvo se odredi broj kodova koji se koristi. On će zavisiti od broja karaktera u tekstu, ali samo do neke granice kada postaje konstantan, jer je statistika karaktera više podložna promeni u kraćim tekstovima, nego u dužim. Zatim se nađe razlika između prvog stepena dvojke koji je veći od broja kodova i broja kodova. Broj koji je stepen dvojke se uzima, jer je kasnije potrebno da se kodovi iz decimalnog oblika prebace u binarni. Razlika predstavlja broj kodova koje treba odbaciti. Mora se voditi računa da broj kodova koji se odabere bude takav da teži stepenu dvojke sa donje strane da bi se manje kodova odbacilo. Kodovi koji su višak se odbacuju na slučajan način i simetrično, ako ih je paran broj, a ako ih je neparan onda se prvo odbaci jedan slučajno. Na primer, ako je mogući broj kodova 16, od 0 do 15, i ako se izbaci kod broj 3 onda se izbaci i kod broj 12 (15-3). U binarnoj predstavi bi to bilo da se izbaci broj sa svojim komplementom: 0011 i 1100. Ovako se formira grupa simbola u kojoj su svi simboli različiti. Ovo će omogućiti da se u binarnoj predstavi jedinice i nule pojave skoro isti broj puta.

Kodovi se sada dodeljuju karakterima tako što svaki karakter dobije određeni broj kodova na sasvim slučajan način i srazmerno svom pojavljivanju u tekstu. Zbog toga ne postoji statistička zavisnost između kodova dodeljenih jednom karakteru. Kada se ovaj proces završi baza kodova je formirana.

Šifrovanje teksta je sledeće što treba uraditi. Sada se karakteri zamenjuju odgovarajućim kodovima iz baze na sasvim slučajan način. Ova operacija ne mora odgovarati redosledu pojavljivanja kodova u bazi, ali se radi tako da se svi kodovi iskoriste isti ili vrlo približni broj puta za odgovarajući karakter. Ovime je narušena i statistička zavisnost između kodova u samom kriptogramu. Na primer, ako se koriste četiri koda za karakter koji se pojavljuje:

- 15 puta – onda će se 3 koda pojaviti po 4 puta, a jedan 3 puta;
- 16 puta – onda će se sva 4 koda pojaviti po 4 puta;
- 17 puta – onda će se 3 koda pojaviti po 4 puta, a jedan 5 puta;
- 18 puta – onda će se 2 koda pojaviti po 4 puta, a dva po 5 puta.

Nakon ovakvog šifrovanja u prvom delu algoritma u drugom delu će se sigurno dobiti kriptogram u kome nema statističke zavisnosti između iskorišćenih karaktera ASCII tabele. Može se desiti da entropija pokaže suprotno. Entropija parova može biti manja nego entropija pojedinačnih karaktera, što znači da u tekstu postoji statistička zavisnost. Ovaj podatak kriptoanalitičarima neće ništa značiti. Oni će misliti da su došli do nekog zaključka, sami će sebe navesti na pogrešan put, a ustvari nisu, nego su se vratili na početak. Zašto, biće objašnjeno u nastavku.

C. Šifrovanje uz kompresiju

Šifra u prethodnom delu je opisana za slučaj kada se šifruju pojedinačni karakteri u poruci. Pored toga može se posmatrati i šifrovanje parova. Parovi se biraju tako da je neophodno šifrovati upola manje karaktera, kada ih je paran broj, jer se vezuju prvi i drugi, treći i četvrti itd. Kada je neparan broj karaktera, doda se jedan na kraj ili na početak pa se onda posmatra kao paran broj. Ovaj metod je pogodan, jer se dve binarne reči od po 8 bita mogu šifrovati sa kodnom reči od, na primer, 12 bita. Tajnost tako dobijenog kriptograma je potpuno ista kao i kada se šifruju pojedinačni karakteri, samo je sada omogućena i kompresija podataka.

Ovaj metod se može primenjivati i kod kraćih i kod dužih tekstova. Kod nekih dugačkih tekstova, kod kojih se javlja dosta različitih kombinacija karaktera neće biti neka velika kompresija, možda je, čak, uopšte i neće biti, ali se veličina kriptograma sigurno neće povećati u odnosu na veličinu poruke, a zaštita će uvek biti dobra.

IV PRIMERI

A. Formiranje kriptograma

U ovom delu je prikazan način na koji se formira baza kodova i kako se od tih kodova formira kriptogram sastavljen od brojeva i kriptogram sastavljen od karaktera iz ASCII tabele. Tekst se sastoji od 225 karaktera i iz ovog primera se ne vidi kvalitet šifre, jer je tekst kratak, već on služi samo da bi se pokazao princip rada ove metode šifrovanja.

Primer:

Nikola Tesla je jedan od najpoznatijih svetskih pronalazaca i naucnika u oblasti fizike, elektrotehnike i radiotehnike. Roden je 10. jula 1856, u selu Smiljan kod Gospica u srpskoj porodici, a umro 7. januara 1943 u Njujorku.

Statistika karaktera u tekstu je data u tabeli 1. U prvoj koloni su dati redni brojevi karaktera iz ASCII tabele koji se pojavljuju u tekstu, a u zagradi se nalaze odgovarajući karakteri. Postoji 38 različitih karaktera. U drugoj koloni su date vrednosti, koje pokazuju koliko se puta koji od njih pojavljuje. U ovoj tabeli je prikazano i koliko se kom karakteru dodeljuje kodnih simbola od ukupno 48 kodova. To je dato u trećoj koloni. Svaki kod se može da pojavi najviše deset puta u kriptogramu, a naravno, mora se pojaviti bar jednom.

TABELA 1: STATISTIKA TEKSTA, BROJ KODOVA I KODOVI SVAKOG KARAKTERA

karakteri (redni broj iz ASCII tabele)	broj pojavljivanja u tekstu	broj kodova	kodovi
32 ( )	35	4	60, 48, 61, 43
44 (,)	3	1	56
46 (.)	4	1	27
48 (0)	1	1	54
49 (1)	3	1	30
51 (3)	1	1	1
52 (4)	1	1	20
53 (5)	1	1	58
54 (6)	1	1	25
55 (7)	1	1	6
56 (8)	1	1	13
57 (9)	1	1	36
71 (G)	1	1	59
78 (N)	2	1	34
82 (R)	1	1	14
83 (S)	1	1	2
84 (T)	1	1	49
97 (a)	20	2	51, 4
98 (b)	1	1	44
99 (c)	4	1	9
100 (d)	6	1	15
101 (e)	14	2	33, 22
102 (f)	1	1	57
104 (h)	4	1	41
105 (i)	17	2	38, 21
106 (j)	11	2	47, 39
107 (k)	10	1	19
108 (l)	8	1	42
109 (m)	2	1	12
110 (n)	11	2	3, 55
111 (o)	15	2	5, 50
112 (p)	5	1	26
114 (r)	8	1	62
115 (s)	8	1	37
116 (t)	6	1	29
117 (u)	11	2	24, 7
118 (v)	1	1	8
122 (z)	3	1	16

U tabeli 1, u četvrtoj koloni se nalaze kodovi koji su dodeljeni odgovarajućim karakterima. Evo i načina na koji su oni izabrani. Pošto se koristi 48 kodova, a prvi naredni broj, koji je stepen dvojke je 64, onda će se od brojeva od 0 do 63 odabrati 48. Kao što je već rečeno, da bi se odredili kodovi mora se proizvoljno odbaciti 16 brojeva, 8 koji će se slučajno odrediti i 8 njima komplementarnih brojeva u binarnoj predstavi u datom intervalu. U tabeli 2 su dati brojevi koji se odbacuju. U prve dve vrste se nalaze njihovi decimalni zapisi, a u naredne dve binarni, jer se tada vidi da su brojevi simetrični i da se uvek izbacuje podjednak broj nula i jedinica.

TABELA 2: ODBAČENI KODOVI

decimalni zapis		binarni zapis	
0	63	000000	111111
10	53	001010	110101
11	52	001011	110100
17	46	010001	101110
18	45	010010	101101
23	40	010111	101000
28	35	011100	100011
31	32	011111	100000

U narednom pasusu je prikazan kriptogram datog teksta nakon primene druge faze algoritma. Kodovi su međusobno razdvojeni razmakom radi preglednosti, a svaki jednocifreni broj je napisan sa prefiksom nula.

Kriptogram:

34 21 19 50 42 04 60 49 22 37 42 51 60 39 33 60 47 33 15 04 55 43 05 15 61 03 04 39 26 05 16 03 04 29 38 47 38 41 61 37 08 22 29 37 19 21 41 43 26 62 50 03 04 42 51 16 04 09 51 43 21 60 03 04 24 09 55 21 19 04 48 24 60 50 44 42 51 37 29 38 43 57 21 16 21 19 22 56 48 33 42 22 19 29 62 50 29 33 41 03 38 19 33 48 38 43 62 04 15 21 05 29 22 41 55 38 19 33 27 61 14 50 15 22 03 60 39 22 43 30 54 27 43 47 24 42 51 61 30 13 58 25 56 48 07 60 37 22 42 24 48 02 12 21 42 39 51 55 43 19 50 15 61 59 05 37 26 21 09 51 48 24 43 37 62 26 37 19 05 39 48 26 50 62 05 15 38 09 38 56 60 51 48 24 12 62 50 48 06 27 60 39 51 55 24 04 62 51 61 30 36 20 01 61 24 61 34 39 07 47 50 62 19 07 27

Sada se svaki kod prebacuje iz decimalnog u svoj binarni oblik dužine 6 bita.

Primenom drugog dela algoritma i deljenjem niza bita na grupe od po 4 bita možemo dobiti kriptogram sastavljen od 16 različitih karaktera. Ovde je uzeto da su to prvih 16 slova engleske abecede. Ovakav kriptogram je prikazan u sledećem pasusu.

Kriptogram:

IJFEPCKIEPDBFKFKLDPCHIHMLOBDMENOLBEPPE DBCHGIFEADBBNJKPJKJPGFCBGHGFENFKGLGLO MIDBCKMNABAJMOLFHMAMEGAJNNFEMEMBIPDC LCKMOFHGGKPJFFAFFDFLIMCBKJGENNPLCHGBK EDJJDIHAJKLPIEDNFBFNFKJNOGEOBGPNDLCDNG APMJNGKNONJLKOPGCKMPNHINOJJODABPMJFGK JIMACDBFKKHMPHKNDMIPPHLBGFGJFCHDMBIK OFPJKJFDBGHMBKMLOBEPJIJJLIPDDMBIDDOMLA BJLPCHMPHGAEPLDPFOJBEAHNGDNIKHBOPMLO EMHGM

B. Provera kvaliteta šifre

Da bi se pokazao kvalitet šifre potrebno je analizirati i uporediti statističke osobine teksta i dobijenog kriptograma.

Korišćeni tekst se sastoji od 8444 karaktera, a od toga je 71 različit karakter (deo teksta sa Internet adrese [3]). U

njemu se, kao što se i očekivalo, najviše pojavljuje razmak, oko 1300 puta, a zatim samoglasnici i to mala slova, svako oko 650 puta. Na osnovu verovatnoće pojavljivanja karaktera izračunata je entropija pojedinačnih karaktera (4.5502) i entropija parova (3.3519) u tekstu [4]. Tako dobijeni rezultati pokazuju da u tekstu postoji statistička zavisnost između karaktera, a ona je direktna posledica samog sadržaja poruke.

Za šifrovanje teksta je potreban 971 kod, odnosno decimalni broj. Srazmerno broju pojavljivanja najviše kodova dobijaju razmak i samoglasnici. U kriptogramu se svaki od ovih kodova može pojaviti najviše 9 puta. Nakon šifrovanja teksta dobija se kriptogram u kome postoji 815 kodova koji se pojavljuju 9 puta i 113 kodova koji se pojavljuju 8 puta. Verovatnoća da se u kriptogramu pojavi neki od ovih kodova je 95,57%. Na ovaj način je postignuto da nema simbola koji se pojavljuju drastično više puta nego neki drugi. Procenat simbola koji se pojavljuju manji broj puta je zanemariv, ali to nije podatak koji je koristan za kriptoanalitičare, jer ti kodovi mogu da budu od nekog karaktera koji se pojavljuje više puta, a i ne moraju.

Kada se svaki decimalni broj zameni njegovom binarnom predstavom dužine 10 bita dobiće se niz nula i jedinica koji se pojavljuju približno isti broj puta (verovatnoća da se pojavi jedinica je 50,008%).

Nakon primene drugog dela algoritma dobija se kriptogram sastavljen od karaktera iz ASCII tabele. Radi analize ovde je urađeno 5 slučajeva, koji se međusobno razlikuju po broju različitih karaktera, 256, 128, 64, 32 ili 16. Njihove entropije i indeksi koincidencije, IC (*Index of Coincidence*) [4], su dati u tabeli 3.

TABELA 3: ENTROPIJE I INDEKSI KOINCIDENCIJE KRIPTOGRAMA

broj različitih karaktera	entropija pojedinačnih karaktera	entropija parova karaktera	indeks koincidencije
256	7.9838	5.2252	0.0039
128	6.993	5.9297	0.0078
64	5.9974	5.7841	0.01561
32	4.9993	4.9666	0.03122
16	3.9999	3.9937	0.06247

Indeks koincidencije određuje stepen varijacije učestalosti pojavljivanja simbola u kriptogramu. Što je perioda šifre veća to je vrednost indeksa koincidencije manja. Iz ove analize se čini da je perioda ponavljanja ključa za šifrovanje veoma duga za sve sem za poslednji kada se predpostavlja da je perioda ustvari jedan. Naravno, ovo će ili zavarati kriptoanalitičare ili neće, ali za dekriptovanje šifre neće pomoći, jer ovde nije korišćena šifra u kojoj postoji neka periodičnost pojave karaktera.

Jedna od karakteristika dobijenih kriptograma je i to da se svi simboli pojavljuju približno isti broj puta i potpuno slučajno. Ipak, entropije za 256 i 128 različitih karaktera to ne pokazuju. Zašto? Kada se simboli ponavljaju manji broj puta, kao što je to ovde slučaj, onda ispada da postoji statistička zavisnost među njima, a nje ustvari nema.

Evo i objašnjenja. Na primer, imamo neki tekst od ukupno 640, a 128 različitih karaktera. Svaki od njih se pojavljuje u proseku po 5 puta sasvim slučajno. U formuli za entropiju figuriše verovatnoća $P(s_j/s_i)$, verovatnoća da se u tekstu posle pojavljivanja simbola s_i pojavi simbol s_j [1]. Povećanje ove verovatnoće utiče na smanjenje entropije kada se posmatraju parovi simbola. Ona je ovde za svaku kombinaciju velika i u proseku iznosi oko 20% (zbir svih ovih verovatnoća podeljen sa brojem ostvarenih kombinacija). Kada bi se 128 karaktera sasvim slučajno i približno isti broj puta rasporedilo u tekst od 150.000 simbola prosečna verovatnoća bi drastično opala i bila bi u proseku 0,78125%, a entropija bi porasla, jer je tada moguće ostvariti svih 16.384 kombinacija. Ovo je moguće ostvariti tek kada je broj simbola u tekstu u proseku 9 puta veći od broja mogućih kombinacija.

Upravo ovaj opisani efekat pomaže da se objasne rezultati prikazani u tabeli 3. Oni navode kriptoanalitičare na pogrešan put, jer statistika pojavljivanja karaktera u kriptogramu ne zavisi od statistike pojavljivanja karaktera u poruci, što je glavna prednost opisane šifre.

Pri šifrovanju parova karaktera kvalitet šifre i rezultati kriptoanalize su maltene isti. Na ovaj način se još može postići i kompresija podataka.

V ZAKLJUČAK

Prednosti opisane šifre su to što je pogodna za zaštitu ličnih podataka koji se ne prenose već se čuvaju na disku i što zadovoljava dva bitna uslova:

- tajnost, tako što za razbijanje šifre nije dovoljno imati kriptogram i znati metod formiranja šifre i dužinu kodne reči i
- autentičnost, jer se ne zna koji je karakter iz poruke zamenjen kojom kodnom reči, pa se zbog toga ne može izmeniti sadržaj kriptograma, a da se to ne primeti.

LITERATURA

[1] Dušan B. Drajić, „Uvod u teoriju informacija i kodovanje“, Akademska Misao, Drugo izdanje, 2004.

[2] Internet adresa: http://www.wikipedia.org, besplatna enciklopedija opšteg sadržaja dostupna samo na Internetu

[3] Internet adresa: http://sr.wikipedia.org/sr-el/%D0%9D%D0%B8%D0%BA%D0%BE%D0%BB%D0%B0_%D0%A2%D0%B5%D1%81%D0%BB%D0%B0

[4] Bruce Schneier, ”Applied Cryptography“, Jonh Wiles & Sons, Second Edition, 1996.

ABSTRACT

In this paper a novel cipher based on statistical characteristics of text is presented. The cipher satisfies all secrecy and authentication requirements. Also, it satisfies one essential characteristic which breaking cipher make more difficult. This characteristic is that the way of appearing symbols in cipher text does not depend on way of appearing in original text.

CIPHER BASED ON STATISTICAL CHARACTERISTICS OF TEXT

Luka Milinković