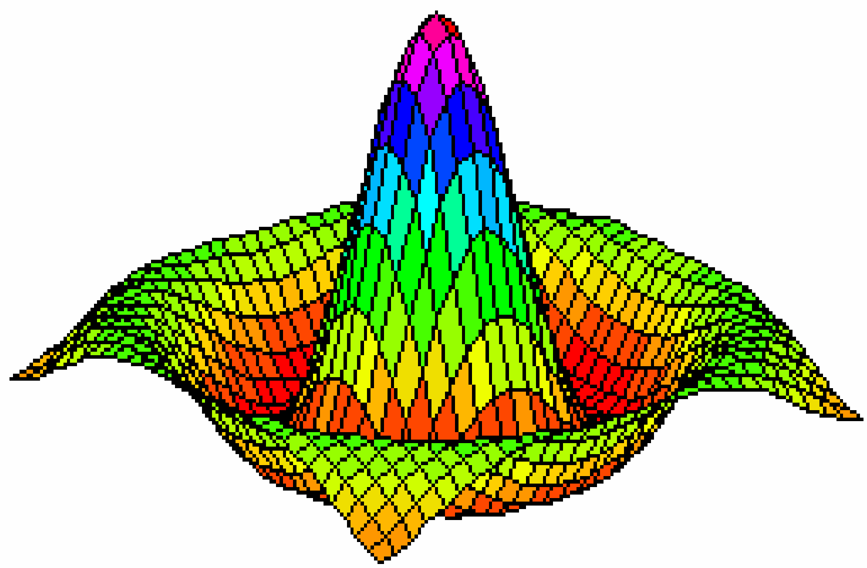
SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA
ZAVOD ZA TELEKOMUNIKACIJE

Matija MIKAC



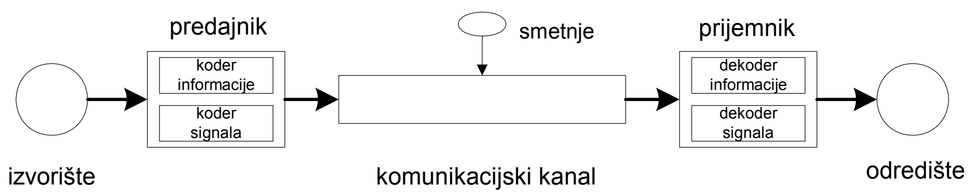
LABORATORIJSKA VJEŽBA

Teorija informacije

ENTROPIJSKO KODIRANJE

Informacijski komunikacijski sustavi

Općeniti informacijski komunikacijski sustav prikazan je na slici 1. Takav sustav uključuje izvorište informacije, predajnik, komunikacijski kanal, prijemnik i odredište informacije. Izvorište generira informacije, te mu stoga pridajemo posebnu pažnju. Svaka generirana informacija sastavljena je od slijeda ulaznih simbola sadržanih u izvorišnoj abecedi. Svaki simbol izvorišne abecede javlja se unutar informacije s određenom učestalošću. Prije prijenosa, informacija se mora kodirati – ulazni simboli prikazuju se korištenjem simbola iz kodne abecede (u našem slučaju, binarni simboli). Budući da ulazni simboli imaju različite frekvencije pojavljivanja, moguće je postupkom kodiranja smanjiti informacijsko opterećenje sustava. Kodovi kojima se služimo kako bi smanjili informacijsko opterećenje nazivaju se **optimalnim kodovima.** Kako optimalni kodovi omogućuju sažimanje informacije nazivamo ih i **kompresijskim kodovima**.



Slika 1 : Informacijski komunikacijski sustav

Spomenuta kodiranja obavljaju se u dijelu predajnika koji se zove koder informacije. Osim kodera informacije predajnik uključuje i koder signala koji kodira prijenosni signal kako bi ga prilagodio prijenosnom mediju. Informacije koje dolaze na odredište potrebno je stoga prije prihvaćanja dekodirati, a to obavljaju dekoder signala i dekoder informacije koji se nalaze u prijemniku.

Metode optimalnog kodiranja

Svaka informacija sastavljena je od slijeda simbola, sadržanih u izvorišnoj abecedi simbola. Općenito, svaki pojedini simbol javlja se sa određenom frekvencijom pojavljivanja u informaciji (npr. unutar riječi/informacije *abrakadabra*, slovo/simbol *a* se javlja pet puta, slovo/simbol *b* dva puta itd.).
Ukoliko ulazne simbole koji se češće javljaju unutar informacije kodiramo koristeći kraći slijed kodnih simbola, a one koji se češće javljaju dužim slijedom, možemo postići smanjenje informacijskog opterećenja – sažimanje ili kompresiju.
Prosječnu duljinu kodne riječi računamo prema izrazu :

$$L = \sum_i p_i \cdot l_i \qquad [bit/simbol] \tag{1}$$

gdje je p_i vjerojatnost pojavljivanja i-tog simbola, a l_i duljina kôdne riječi kojom se prikazuje i-ti simbol. Jedinica je *bit/simbol*. Ukoliko se koriste druge jedinice (npr. *dibit*...) u izrazu se duljina riječi i_i iskazuje u tim jedinicama, a rezultat je u broju tih jedinica po simbolu.
Postavlja se pitanje do koje mjere je moguće komprimirati zapis ulazne abecede. Dakle, potrebno je uvesti mjeru koja će nam pomoći u određivanju količine informacije koju šaljemo. Zbog toga

definiramo entropiju sustava. Entropija ovisi isključivo o vjerojatnostima pojavljivanja simbola unutar informacije. Definirana je kao :

$$H(p) = -\sum_i p_i \log_2 (p_i) \quad [bit/simbol] \tag{2}$$

Najmanja entropija $H(p)$=0 postiže se u sustavu u kojem je vjerojatnost pojave simbola jednaka 1. Budući da je u tom slučaju pojava simbola siguran događaj, ne primamo novu količinu informacije, te je entropija 0. Maksimalna entropija sustava koji generira n simbola je $H(p)=\log_2 n$, To je slučaj kada svi simboli nastupaju sa istom frekvencijom pojavljivanja.

Shannonov teorem o kodiranju u kanalu bez šuma daje slijedeći odnos između srednje prosječne duljine kodne riječi nakon optimalnog kodiranja i entropije sustava.

$$H(p) \le L \le H(p) + 1 \tag{3}$$

Ukoliko je kodiranje ispravno, srednja duljina kodne riječi treba zadovoljavati (3).

Optimalnim Huffmanovim kodiranjem dobivamo neravnomjerni kôd (različite duljine kodnih riječi). Huffmanovi kodovi za simbole veće učestalosti koriste kraće duljine kodnih riječi, a za manje učestale simbole dulje kodne riječi. Dvije najpoznatije metode za određivanje optimalnog koda su Shannon-Fanova metoda i Huffmanova metoda.

Shannon-Fanova metoda određivanja optimalnog koda
[Shannon & Weaver 1949., Fano 1949.]

Ulazne simbole u_i poredamo po padajućim vjerojatnostima. Takav slijed podijelimo u dvije grupe, tako da zbroj vjerojatnosti po grupama bude približno 0,5. Svim simbolima u prvoj grupi dodijelimo oznaku 0, a u drugoj grupi 1. Svaka grupa se dalje dijeli u dvije podgrupe, tako da zbrojevi vjerojatnosti u podgrupama budu približno jednaki – prva podgrupa dobiva binarnu oznaku 0, a druga oznaku 1. Postupak se po grupama ponavlja sve dok u podgrupama ne ostane po jedan simbol.

Primjer 1.:

ulazni simbol u_i	vjero-jatnost p_i	binarni simboli					kodna riječ	duljina kodne riječi l_i
u_1	1/2	0					0	1
u_2	1/4	1	0				10	2
u_3	1/8	1	1	0			110	3
u_4	1/16	1	1	1	0		1110	4
u_5	1/32	1	1	1	1	0	11110	5
u_6	1/32	1	1	1	1	1	11111	5

Prosječna duljina kodne riječi :

$$L = 1\cdot\frac{1}{2} + 2\cdot\frac{1}{4} + 3\cdot\frac{1}{8} + 4\cdot\frac{1}{16} + 5\cdot\frac{1}{32} + 5\cdot\frac{1}{32} = \frac{62}{32} = 1{,}9375 \ \ bit/simbol$$

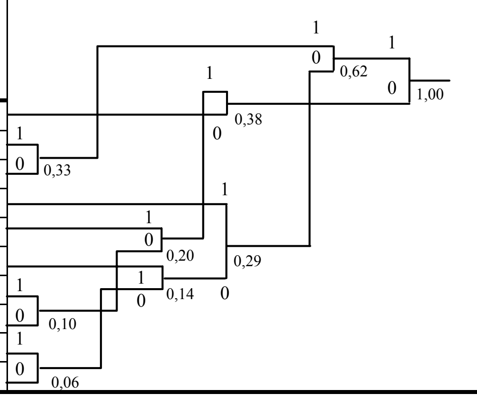
Entropija : $H(p) = L$

Zaključujemo da je kôd koji smo odredili optimalan.

Huffmanova metoda određivanja optimalnog koda
[Huffman 1952.]

Ulazne simbole u_i poredamo po padajućim vjerojatnostima. Dva simbola s najmanjim vjerojatnostima spajamo i zbrajamo njihove vjerojatnosti. Takav spoj simbola postaje dio skupa. Postupak se uvijek nastavlja spajanjem dvaju simbola ili spoja simbola s najmanjim vjerojatnostima, sve dok zbroj vjerojatnosti ne bude 1. Kôd se formira pridruživanjem binarnih oznaka 1 i 0, od korijena (zbroj vjerojatnosti 1) prema vrhovima stabla, tako da oznaku 1 pridružimo gornjem, a oznaku 0 donjem simbolu.

Primjer 2.:



ulazni simbol u_i	vjerojatnost p_i	kodna riječ	duljina kodne riječi l_i
u_1	0,18	00	2
u_2	0,17	111	3
u_3	0,16	110	3
u_4	0,15	101	3
u_5	0,10	011	3
u_6	0,08	1001	4
u_7	0,05	0101	4
u_8	0,05	0100	4
u_9	0,04	10001	5
u_{10}	0,02	10000	5

Srednja duljina kodne riječi :

$$L = 2 \cdot 0{,}18 + 3 \cdot 0{,}17 + 3 \cdot 0{,}16 + 3 \cdot 0{,}15 + 3 \cdot 0{,}10 + 4 \cdot 0{,}08 + 4 \cdot 0{,}05 + 4 \cdot 0{,}05 + 5 \cdot 0{,}04 + 5 \cdot 0{,}02 =$$
$$= 3{,}12 \quad bit / simbol$$

Entropija izvora :

$$H(p) = 3{,}07 \ \ bit / simbol$$

Vrijedi $3{,}07 < 3{,}12 < 4{,}07$, pa zaključujemo da je dobiveni kôd optimalan.

Shannon-Fanova metoda ne garantira formiranje optimalnog Huffmanovog koda. Razlike rezultata kodiranja korištenjem obje metode prikazane su u sljedećem primjeru. Provjerite!

Primjer 3.:

ulazni simbol u_i	vjero-jatnost p_i	binarni simboli			kodna riječ	duljina kodne riječi l_i
u_1	0,35	0	0		00	2
u_2	0,17	0	1		01	2
u_3	0,17	1	0		10	2
u_4	0,16	1	1	0	110	3
u_5	0,15	1	1	1	111	3

$$\overline{b} = 0{,}70 + 0{,}34 + 0{,}34 + 0{,}48 + 0{,}45 = 2{,}31 \ bit/simbol$$

<table>
<tr><th>ulazni simbol
u_i</th><th>vjero-jatnost
p_i</th><th colspan="8">binarni simboli</th><th>kodna riječ</th><th>duljina kodne riječi
l_i</th></tr>
<tr><td>u_1</td><td>0,35</td><td colspan="7">1</td><td rowspan="5">1,00</td><td>1</td><td>1</td></tr>
<tr><td>u_2</td><td>0,17</td><td colspan="3">1</td><td rowspan="2">0,34</td><td rowspan="2">1</td><td rowspan="4">0,65</td><td rowspan="4">0</td><td>011</td><td>3</td></tr>
<tr><td>u_3</td><td>0,17</td><td colspan="3">0</td><td>010</td><td>3</td></tr>
<tr><td>u_4</td><td>0,16</td><td>1</td><td rowspan="2">0,31</td><td colspan="2" rowspan="2"></td><td rowspan="2">0</td><td>001</td><td>3</td></tr>
<tr><td>u_5</td><td>0,15</td><td>0</td><td>000</td><td>3</td></tr>
</table>

$$L = 0{,}35 + 0{,}51 + 0{,}51 + 0{,}48 + 0{,}45 = 2{,}30 \ bit/simbol$$

Entropija izvora je *H(p)* = 2,23 *bit/simbol*. Jasno je da oba dobivena koda zadovoljavaju (3). Međutim, kôd koji je dobiven Huffmanovom metodom daje kraću prosječnu kodnu riječ.

Opisani optimalni kodovi predstavljaju osnovu za kompresiju informacija bez gubitaka. Primjena je vrlo široka, a vjerojatno najpoznatiji primjer je ZIP algoritam za kompresiju, koji Huffmanovu metodu kombinira sa RLE (*Run Length Encoding*) metodom kompresije. Slično funkcioniraju i ostali alati za sažimanje. Jedan od najboljih algoritama je RAR, koji se koristi u istoimenom alatu. Kombiniranjem različitih kompresijskih metoda dobivaju se izuzetni rezultati. RLE kompresija svodi se na pronalaženje ponavljajućih informacija, te definiranje tablica koje grupu podataka predstavljaju jednostavnijim (kraćim) zapisom.

ZADACI ZA PRIPREMU:

Optimalno kodiranje

O-1. Kodirajte optimalnim binarnim kôdom, korištenjem Huffmanove i Shannon-Fanove metode, informaciju: "Laboratorijske vježbe iz kolegija TINF". Razlikujte mala i velika slova u informaciji. Ponovite postupak kodiranja, ali bez razlikovanja malih i velikih slova!

Komunikacijski sustavi

K-1. Objasnite funkcioniranje komunikacijskog sustava – zbog čega se koristi optimalno kodiranje, zbog čega zaštitno? Objasnite utjecaj smetnji na informaciju koja se prenosi. Na primjeru prikažite mogućnosti neispravnog dekodiranja informacije na odredištu.

ZADACI – Matlab

MO-1. Napišite Matlab funkciju za kodiranje i dekodiranje optimalnim binarnim kodom, korištenjem Huffmanovog postupka kodiranja. Pretpostavite izvor koji generira 5 vijesti – omogućite unos vjerojatnosti pojavljivanja pojedine vijesti. Izračunajte entropiju H i srednju duljinu kodne riječi L.

MO-2. Napišite Matlab funkciju za kodiranje i dekodiranje optimalnim binarnim kodom, korištenjem Shannon-Fanovog postupka kodiranja. Pretpostavite izvor koji generira 5 vijesti – omogućite unos vjerojatnosti pojavljivanja pojedine vijesti. Izračunajte entropiju H i srednju duljinu kodne riječi L.

MO-3. Kao i MO-1, no izvor definirajte unesenom informacijom koju treba kodirati (npr. rečenica - ograničite duljinu informacije na 15 simbola).

MO-4. Kao i MO-2, no izvor definirajte unesenom informacijom koju treba kodirati (npr. rečenica - ograničite duljinu informacije na 15 simbola).