EFIKASNO KODOVANJE INFORMACIJA

Uvod

Univerzitet Sinergija

Ciljevi lekcije

- Osnove kodovanja diskretnog izvora informacija u cilju kompresije poruka
- Koje uslove mora da zadovolje kodovi u cilju efikasne kompresije
- Fundamentalna granica kompresije podataka
- Metode konstrukcije efikasnih kompresionih kodova

Uvod

- Prvi praktični problem koji je moguće rešavati zahvaljujući Šenonovom konceptu mera neodredjenosti jeste efikasno kodovanje diskretnog izvora informacija u formi niza simbola.
- Pokazaće se da je entropija izvora fundamentalno ograničenje za kompresiju podataka, odnosno da ona odredjuje minimalnu očekivanu vrednost dužine koda.
- Ali, pre svega razmotrimo pitanje:
- Šta je razlog da uopšte nešto kodujemo?

Uvod

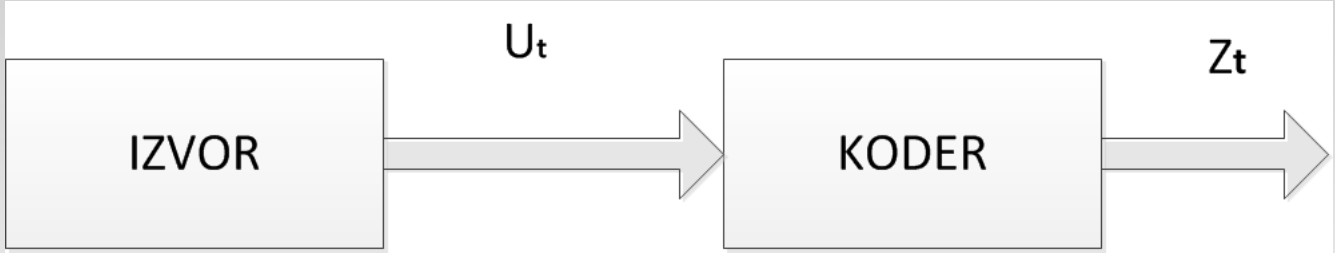
- Postoje tri osnovna razloga za to:

1. Kodovanje radi kompresije (izbacivanje redundanse)
2. Kodovanje u cilju kvalitetnog prenosa u uslovima šuma na kanalu (dodavanje redundanse)
3. Kodovanje radi tajnosti - poruke su dostupme samo autorizovanim korisnicima

- Ova lekcija će se baviti prvim aspektom efikasnog kodovanja diskretnog izvora informacija

Kodovanje jedne slučajne veličine

- Izvor informacija je generator poruka odnosno sekvence simbola.



Osnovna šema kodovanja izvora (source coding): simbol izvora U_t u trenutku t se transformiše u odgovarajuću kodnu reč Z_t

Kodovanje jedne slučajne veličine

- Simboli su elementi skupa koga nazivamo alfabet.
- Informacioni izvor je diskretan ako je alfabet konačan.
- Dimenzija alfabeta se naziva i arnost izvora.
- Razmatraćemo samo poruke konačne dužine.

Primer:

- Novine se mogu smatrati jednim diskretnim izvorom informacija, nad alfabetom datog jezika, dok su poruke novinski tekstovi.

Kodovanje jedne slučajne veličine

- Koder transformiše poruke izvora u kodne reči.
- Kodne reči su sekvence simbola iz kodnog alfabeta, koji je u opštem slučaju različit od alfabeta poruka.
- Formalno, informacioni izvor U_t , odnosno U bez indeksa t ako vremenski indeks nije relevantan, je u stavri slučajan proces nad datim alfabetom ν_U. Jednostavnije rečeno, to je niz slučajnih veličina iz ν_U.
- Izvor emituje simbole u svakom trenutku t sa verovatnoćama $P(U_t = u_i)$.
- Izvor je bez memorije, ukoliko verovatnoća simbola u_i ne zavisi od prošlih emitovanih vrednosti

$$\forall t \geq 1 \; \forall u_i \in \mathcal{V}_U \quad P(U_t = u_i | U_1 ... U_{t-1}) = P(U_t = u_i).$$

Kodovanje jedne slučajne veličine

Univerzitet Sinergija

- Izvor informacija je stacionaran, ukoliko verovatnoće emitovanja simbola $P(U_t = u_i)$ ne zavise od vremena t.
- Koder možemo posmatrati kao preslikavanje $Z = f(U)$ iz alfabeta izvora ν_U u skup kodnih reči ν_Z.
- Ako sa Z^*označimo skup svih konačnih sekvenci sastavljenih od simbola iz kodnog alfabeta ν_Z, tada su kodne reči zadatog kodera jedan podskup skupa Z^*.

Definicija 2.1 ***Nesingularni kodovi***

Kod jednog diskretnog informacionog izvora se naziva nesingularnim, kada se različitim simbolima izvora korespondiraju različite kodne reči. Formalno,

$$u_i \neq u_j \implies z_i \neq z_j.$$

pri čemu je z_i kodna reč koja korespondira simbolu izvora u_i.

Kodovanje jedne slučajne veličine

- U ovoj lekciji bavićemo se samo nesingularnim kodovima.

Primer 2.1 *Kodovanje konačnog izvora*

Morzeov kod služi za kodovanje slovnih i brojčanih znakova. Alfabet morzeovog koda čine samo dva simbola: tačka (·) i crta (-). Npr. A, E i K se koduju sa "· ", "· -" i "- · -" , respektivno.

- Budući da je cilj kodera kodovanje poruka, a ne pojedinačnih simbola, od interesa su kodovi koji se mogu na jedinstven način dekodovati. Takvi kodovi se zovu nedvosmisleni ili jednoznačno dekodivi kodovi. Kod ovih kodova bilo koja sekvenca kodnih reči korespondira jednoznačno samo jednoj poruci informacionog izvora.

Kodovanje jedne slučajne veličine

Definicija 2.2 ***Jednoznačno dekodivi kodovi***

Kod diskretnog izvora informacija se naziva jednoznačno dekodivim, ako i samo ako bilo koja konačna sekvenca kodnih reči jednoznačno odgovara jednoj poruci izvora.

Primer 2.2 *Nejednoznačan kod*

Neka se izvor sastoji od tri simbola a, b i c. Poruka ovog izvora je svaki niz koji se sastoji od ova tri simbola, npr aaabcaacc.

Kod definisan preslikavanjem

$$\mathtt{a} \mapsto 1 \qquad \mathtt{b} \mapsto 00 \qquad \mathtt{c} \mapsto 11$$

je nejednoznačan. Npr. prilikom kodovanja niza aaaa, dobiće se kodna reč 1111, što je isto kao i da kodujemo niz cc. Ne postoji način prilikom dekodovanja koji će biti u stanju da razlikuje šta je kodovana poruka: aaaa ili cc.

Kodovanje jedne slučajne veličine

Primer 2.3 *Jednoznačno dekodivi kod*

Neka je izvor informacija kao u prethodnom primeru. Neka je kod definisan sa

$$\mathrm{a} \mapsto 1 \qquad \mathrm{b} \mapsto 00 \qquad \mathrm{c} \mapsto 10$$

Može se dokazati da je ovaj kod jednoznačan. Npr. niz 10000 se dekoduje u niz abb, a niz 1000 u cb.

Medju jednoznačnim kodovima od posebnog značaja su tzv. prefiksni kodovi (Prefix – Free Code).

Jedan niz *z* dužine n (n≥ 1) je prefiks nekog drugog niza z' ako i samo ako je prvih n simbola u z' jednako egzaktno nizu *z*.

Npr. abba je prefiks niza abbabc.

Kodovanje jedne slučajne veličine

Definicija 2.3 *Prefiksni kodovi*

Za jedan kod se kaže da je prefiksan ako u njemu ni jedna kodna reč nije prefiks nekoj dugoj kodnoj reči.

Primer 2.4 *Prefiksni kod*

Neka je dat izvor sa tri simbola a,b i c. Za ovaj izvor primer jednog prefiksnog koda je,

$$\mathtt{a} \mapsto 0 \qquad \mathtt{b} \mapsto 10 \qquad \mathtt{c} \mapsto 11$$

dok kod

$$\mathtt{a} \mapsto 1 \qquad \mathtt{b} \mapsto 00 \qquad \mathtt{c} \mapsto 10$$

nije prefiksan, budući da je 1 kod za simbol a, istovremeno i prefiks u nizu 10, koji predstavlja kodnu reči za simbol c.

Kodovanje jedne slučajne veličine

- Odakle potiče naš interes za prefiksnim kodovima?
- Odgovor leži u dve osobine, 2.1 i 2.2.

Osobina 2.1

Svaki prefiksni kod je jednoznačno dekodiv.

- Važno je napomenuti da postoje jednoznačni kodovi koji nisu prefiksni, kao što nam to pokazuje Primer 2.3.

Definicija 2.4

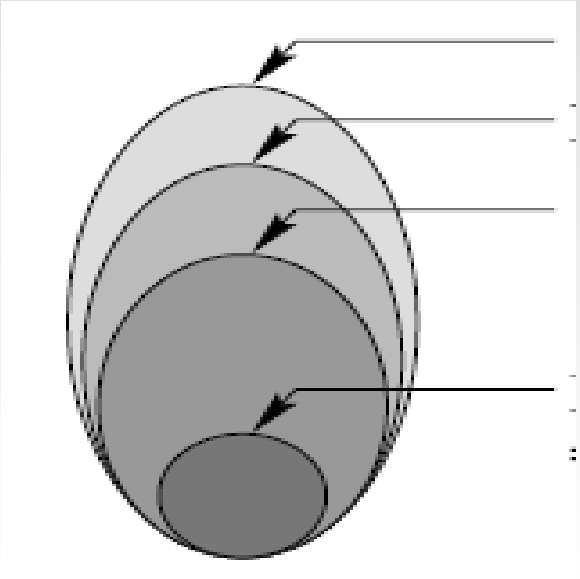
Kod je trenutno dekodiv ako i samo ako se svaka kodna reč u bilo kom nizu kodnih reči može dekodovati u istom momentu kada je dostignut kraj kodne reči.

Osobina 2.2

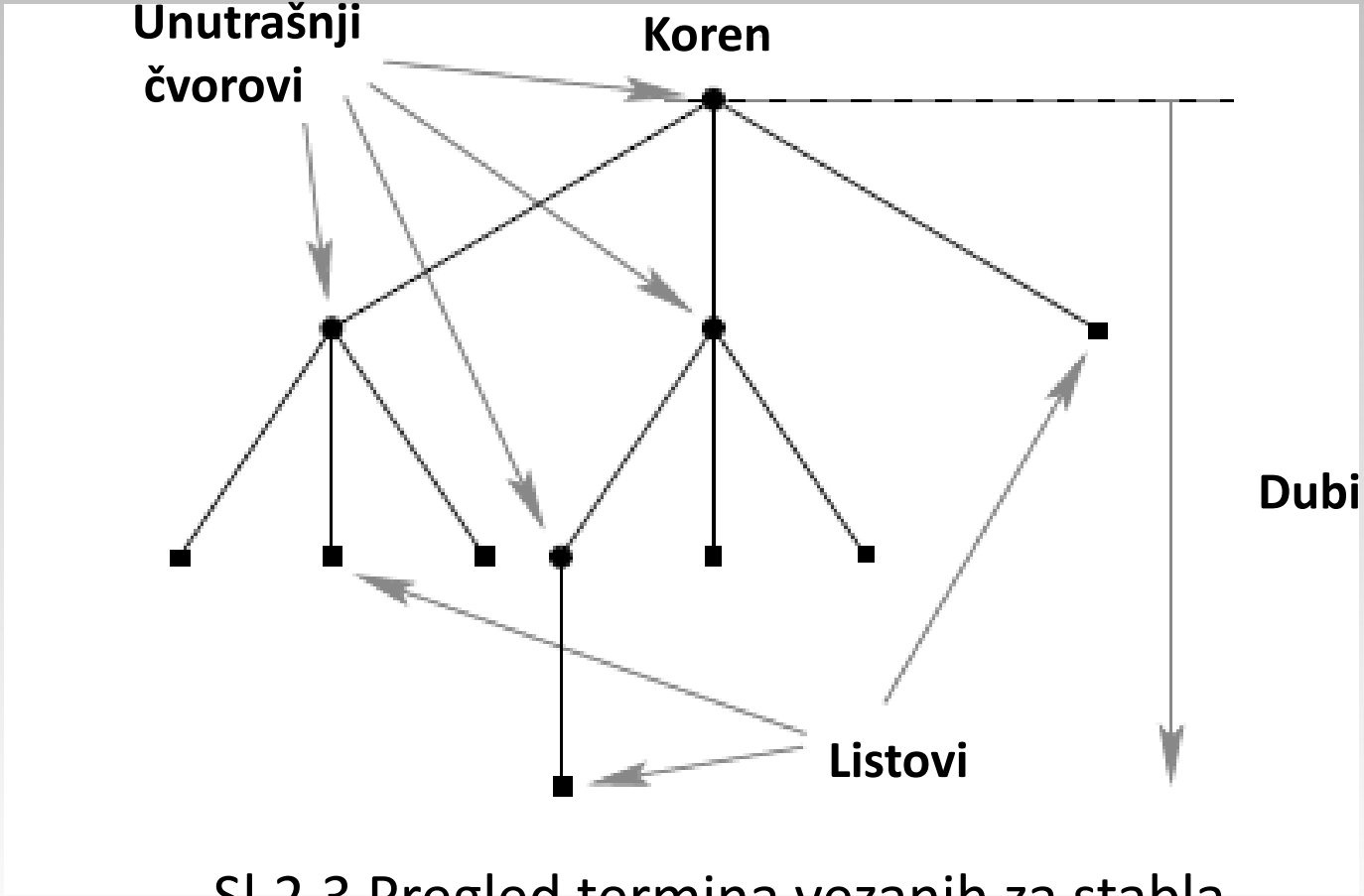
Kod je trenutno dekodiv ako i samo ako je prefiksan.

Kodovanje jedne slučajne veličine

- Trenutno dekodivi kodovi ne zahtevaju memorisanje prispelih kodnih reči ili čekanje dospeća novih da bi dekodovanje bilo obavljeno. Ovim se štedi i vreme i memorijski prostor dekodera.
- Klase kodova koje smo do sada opisali se medjusobno odnose kao na sl.2.2
-



N-arna stabla kodova



Sl.2.3 Pregled termina vezanih za stabla

N-arna stabla kodova

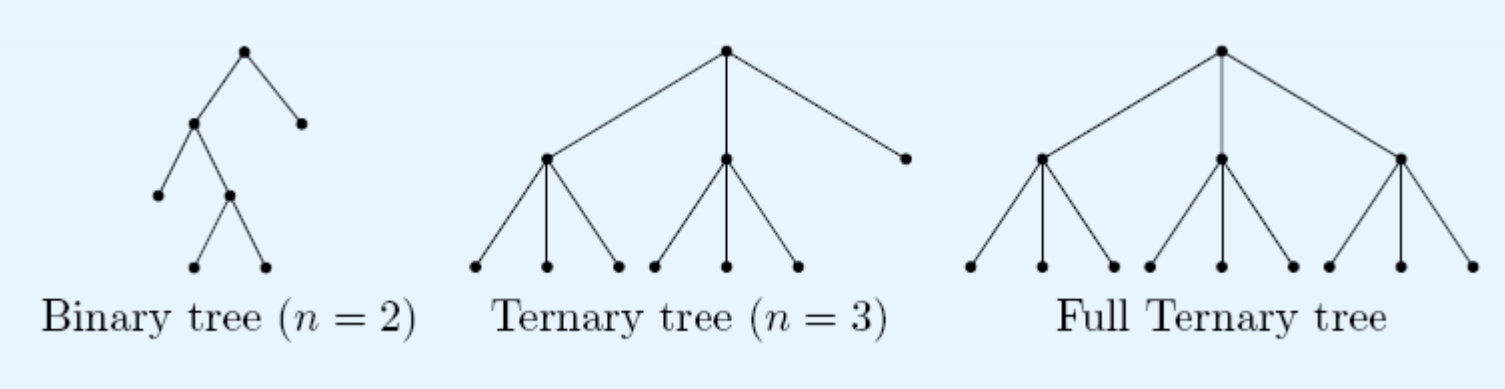
- Pogodan način predstave i izučavanja prefiksnih kodova su n-arna stabla.

Definicija 2.5 *n-arno stablo*

N-arno stablo je stablo u kome svaki unutrašnji čvor ima arnost n, odnosno ima egzaktno n čvorova-dece.

Potpuno n-arno stablo je n-arno stablo u kome svi listovi imaju istu dubinu.

Primer 2.5 *n-arna stabla*



N-arna stabla kodova

- Čvor n_1 <u>pokriva</u> čvor n_2, ukoliko put od korena do n_2 sadrži čvor n_1.

Osobina 2.3

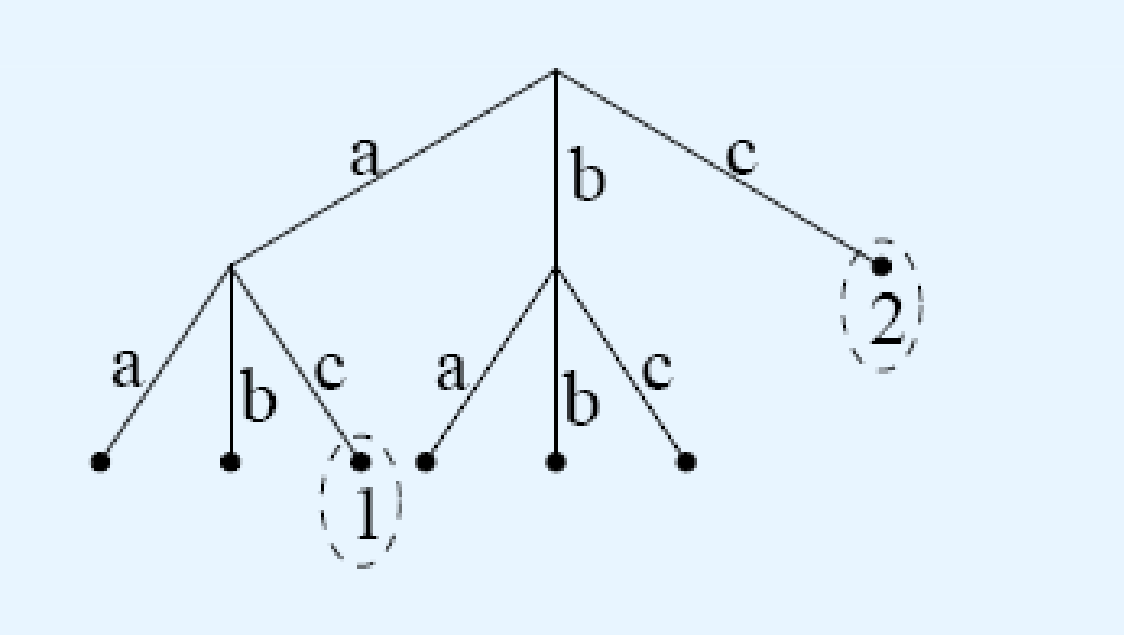
U potpunom n-arnom stablu dubine d≥ $0, \text{svaki čvor dubine } \delta\ (0 \leq \delta \leq d)$ pokriva tačno $n^{d-\delta}$ listova.

Definicija 2.6 *Kodno stablo*

Kodno stablo je n-arno stablo čije su grane označene slovima datog alfabeta dimenzije n, na takav način da se svako slovo pojavljuje samo jedanput u izlaznim granama čvorova. Kodne reči odgovaraju sekvencama oznaka grana duž puteva od korena do listova.

N-arna stabla kodova

Primer 2.6 *Kodno stablo*



Ternarno kodno stablo: kodna reč pretstavljena listom 1 je ac, dok list 2 predstavlja kodnu reč c.

N-arna stabla kodova

Definicija 2.7 *n-arni kod*

Kod čiji alfabet ima n slova se naziva n-arni kod.

Osobina 2.4

Za svaki n-arni prefiksni kod postoji barem jedno n-arno kodno stablo takvo da svaka kodna reč odgovara sekvenci oznaka na jedinstvenom putu od korena stabla do lista.

I obrnuto, svako kodno stablo definiše jedan prefiksni kod. Kodne reči ovog prefiksnog koda su odredjene sekvencom oznaka putanja od korena do svakog lista kodnog stabla.

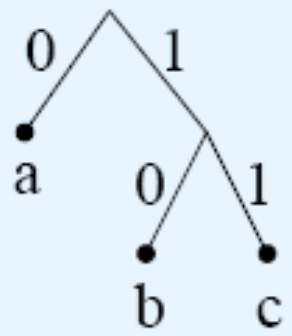
- Ukratko, kodno stablo i prefiksni kod su ekvivalentni objekti.

N-arna stabla kodova

Primer 2.7

Kodno stablo koje odgovara kodu iz primera 2.4

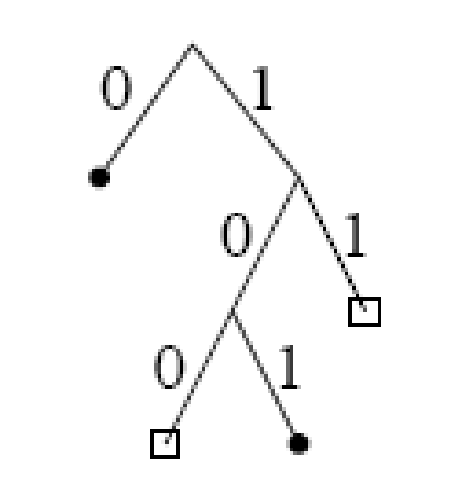
$\mathtt{a} \mapsto 0 \qquad \mathtt{b} \mapsto 10 \qquad \mathtt{c} \mapsto 11$



Prema standardnoj konvenciji prikazivanja, listovi su označeni simbolima izvora, a kodne reči se dobijaju kao nizovi oznaka grana na putanjama od korena do listova.

N-arna stabla kodova

- Primetimo da reprezentacijom prefiksnih kodova pomoću kodnih stabala, neki listovi ne korespondiraju ni jednoj kodnoj reči. Takvi listovi se nazivaju <u>nekorišćeni listovi</u>.
- Primer:
- Binarno kodno stablo koje odgovara kodu $\{0,101\}$ ima dva neiskorićena lista



N-arna stabla kodova

- Zaista nizovi 11 i 100 ne korespondiraju ni jednoj kodnoj reči posmatranog koda.

Definicija 2.8 *Kompletni kod*

Kada nema nekorišćenih listova u odgovarajućem n-arnom kodnom stablu, kod se naziva *kompletni kod*.

Kraftova nejednakost

- Sada su nam potrebni uslovi koji mora da budu ispunjeni da bi jedan prefiksni kod postojao. Ispostavlja se da je Kraftova nejednakost potreban i dovoljan uslov.

TEOREMA 2.1 Kraftova nejednakost

Za svaki D-arni prefiksni kod od N kodnih reči čije su dužine pozitivni celi brojevi $\{l_1, l_2, \dots, l_N\}$, mora važiti

$$\sum_{i=1}^{N} D^{-l_i} \leq 1. \qquad (2.1)$$

i obrnuto, ako su date dužine kodnih reči koje zadovoljavaju ovu nejednakost, postoji prefiksni kod sa kodnim rečima date dužine.

Kada u (2.1) važi jednakost, tada je kod kompletan.

Kraftova nejednakost

- Klasična pogrešna upotreba Kraftove nejednaosti je odgovaranje na pitanje da li je kod sa zadatim dužinama kodnih reči prefiksan. Ova nejednakost nam samo daje informaciju kada može postojati jedan prefiksni kod.
- Teorema 2.1 nam samo kaže da postoji prefiksni kod istih dužina kodnih reči, i zaista on postoji i dat je u istom primeru 2.4, tj kod {0,10,11} .

Kraftova nejednakost

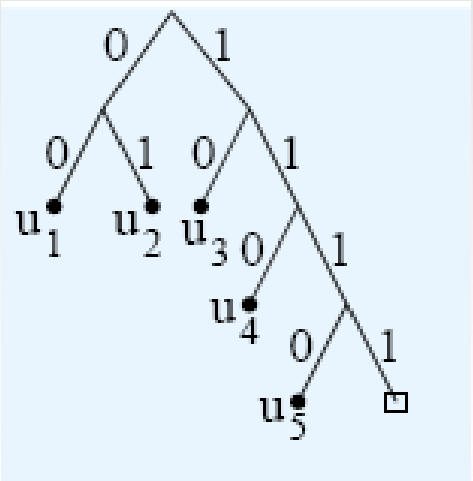
Primer 2.9

Da li postoji binarni prefiksni kod sa dužinama kodnih reči $l_1 = 2$, $l_2 = 2$, $l_3 = 2$, $l_4 = 3$ i $l_5 = 4$?

Odgovor je da , pošto je

$$\sum_{i=1}^{5} 2^{-l_i} = 1/4 + 1/4 + 1/4 + 1/8 + 1/16 = 15/16 < 1.$$

Primer jednog takvog koda je



U	u_1	u_2	u_3	u_4	u_5
Z	00	01	10	110	1110

Kraftova nejednakost

Primer 2.10

Da li binarni prefiksni kod sa dužinama kodnih reči 1,2,2,3,4 postoji? Odgovor je ne, budući da važi

$$\sum_{i=1}^{5} 2^{-l_i} = 1/2 + 1/4 + 1/4 + 1/8 + 1/16 = 19/16 > 1.$$

Efikasno kodovanje

- Osnovni cilj efikasnog kodovanja je kodovanje datog izvora informacija tako da je srednja dužina koda minimalna.
- Pod vrlo opštim uslovima na izvor informacija, koji su skoro uvek ispunjeni u praksi, minimizacija srednje dužine koda je ekvivalentno minimizaciji očekivane dužine koda.

Efikasno kodovanje

Definicija 2.9 *Očekivana dužina koda*

Neka je dat izvor, čiji se simboli u_i emituju sa verovatnoćama p_i, $(0 \leq i \leq N)$ i neka su $l_i$dužine kodnih reči. Očekivana dužina koda, odnosno očekivana dužina kodnih reči je data sa

$$E[L] = \sum_{i=1}^{N} p_i \, l_i$$

Naš cilj je da pronadjemo prefiksni kod koji ima E[L] što je moguće manje.

Jasno nam je da kraćim kodnim rečima treba kodovati najverovatnije simbole. Kako odrediti dužine kodnih reči? Koja je najmanja teorjska vrednost za E[L]?

Verovatnosno n-arno stablo

- Prefiksni kodovi se mogu definisati n-arnim stablima u kojima svaka kodna reč odgovara jednom listu stabla, odnosno jedinstvenom putu od korena do tog lista.
- Budući da izvor emituje simbole sa odredjenim verovatnoćama, jasno je da će i svaka kodna reč, odnosno list, imati tačno odredjenu verovatnoću. Neiskorišćenim listovima se po konvenciji dodeljuju verovatnoće jednake 0.
- Nakon označavanja listova odgovarajućim verovatnoćama, možemo sračunati i verovatnoće unutrašnjih čvorova, rekurzivno od listova ka korenu, imajući uvek u vidu da je verovatnoća jednog čvora jednaka sumi verovatnoća njegivih deca-čvorova.
- Nakon ove procedure dibili smo verovatnosno n-arno stablo.

Verovatnosno n-arno stablo

Definicija 2.10 *Verovatnosno n-arno stablo*

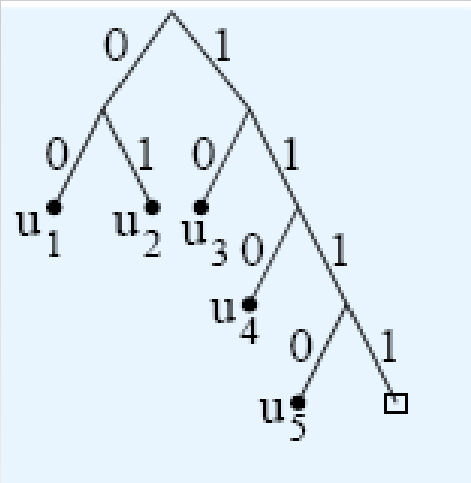
Verovatnosno n-arno stablo je n-arno stablo čijim čvorovima su dodeljene verovatnoće na sledeći načim:

1. Korenu je dodeljena verovatnoća 1.
2. Verovatnoća svih ostalih čvorova, uključujući i koren, jednaka je sumi verovatnoća njegovih deca-čvorova.

Verovatnosno n-arno stablo

Primer 2.11

Za binarni prefiksni kod iz primera 2.9, data sa



U	u_1	u_2	u_3	u_4	u_5
Z	00	01	10	110	1110

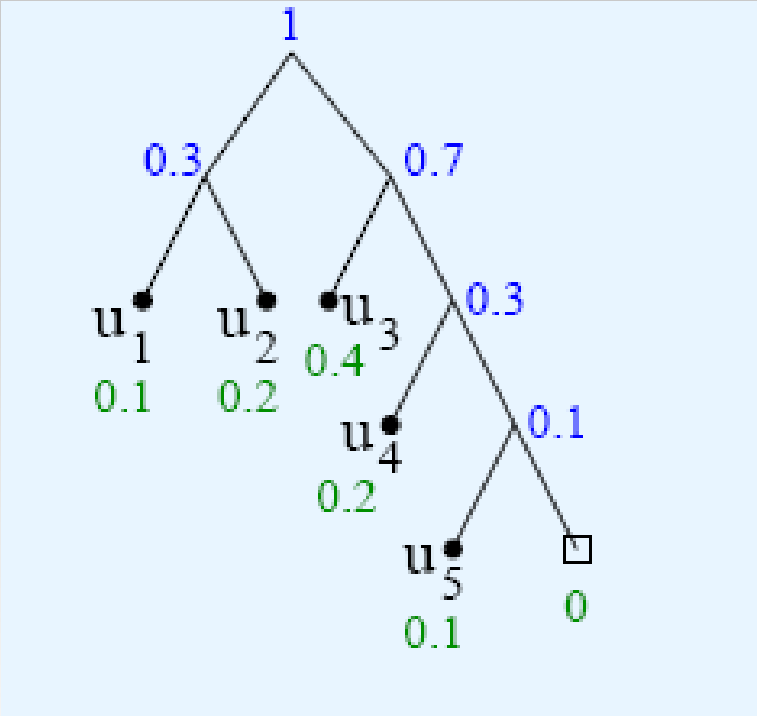
Neka su verovatnoće kodnih reči

$$p_1 = p_5 = 0.1,\ p_2 = p_4 = 0.2 \qquad p_3 = 0.4$$

Tada kao rezultat dobijamo verovatnosno binarno stablo

Verovatnosno n-arno stablo

Nastavak primera 2.11



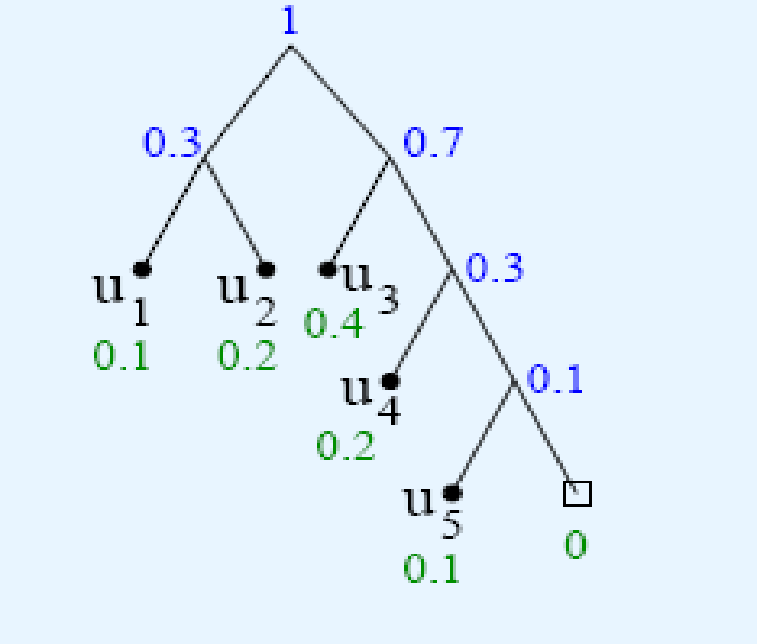
Primetimo da u verovatnosnom n-arnom stablu suma verovatnoća svih listova mora biti jednaka 1.

Verovatnosno n-arno stablo

Lema 2.1 *Lema dužine puta*

U verovatnosnom n-arnom stablu, prosečna (očekivana) vrednost dubine listova jednaka je sumi verovatnoća unutrašnjih čvorova (isključivanjem listova i uključivanjem korena).

Primer 2.12 *Lema dužine puta*



Za primer ovog verovatnosnog binarnog stabla,
očekivana dubina listova je jednaka
1+0.3+0.7+0.3+0.1=2.4
Provera:
očekivana dužina kodnih reči je
2·0.1+2·0.2+2·0.4+3·0.2+4·0.1=2.4

Verovatnosno n-arno stablo

Definicija 2.11 *Entropija listova n-arnog verovatnosnog stabla*

Neka je N broj listova verovatnosnog n-arnog stabla i neka su $p_1, p_2, \dots, p_N$ njihove verovatnoće. Pod entropijom listova podrazumevaćemo entropiju definisanu sa

$$H_{\text{leaf}} = -\sum_i p_i \log p_i$$

Osobina 2.5

Za verovatnosno n-arno stablo koje odgovara prefiksnom kodu informacionog izvora U, važi

$$H_{leaf} = H(U)$$

Verovatnosno n-arno stablo

Dokaz osobine 2.4

Neka je Z prefiksni kod izvora U. Po definiciji verovatnoća p_i je verovatnoća i-te kodne reči, pa je stoga $H_{leaf} = H(Z)$. S druge strane budući da je kod nesingularan, odnosno Z=f(U) je injektivno preslikavanje, H(Z)=H(U). Stoga je H_{leaf}=H(U).

Definicija 2.12 *Entropija grananja*

Neka je M broj unutrašnjih čvorova verovatnosnog n-stabla i neka su $P_1, P_2, \ldots P_M$ njihove verovatnoće. Neka su $q_{i1}, q_{i2}, \ldots, q_{in_i}$ verovatnoće n_i čvorova-dece (uključujući i listove) unutrašnjeg čvora čija je verovatnoća P_i. Entropija grananja H_i u tom čvoru je data sa

$$H_i = -\sum_{j=1}^{n_i} \frac{q_{ij}}{P_i} \log \frac{q_{ij}}{P_i}, \quad P_i = \sum_{j=1}^{n_i} q_{ij}.$$

Verovatnosno n-arno stablo

Primer 2.13

Neka je M=5 čvorova verovatnosnog stabla iz primera 2.9 tako numerisano da je P_1=1, P_2=0.3, P_3=0.7, P_4=0.3, P_5=0.1. Tada je

$$H_{\text{leaf}} = -\sum_{i=1}^{5} p_i \log p_i \simeq 2.122 \text{ bit.}$$

$$n_1 = 2 \qquad q_{11} = 0.3 \qquad q_{12} = 0.7,$$

$$H_1 = -0.3 \log 0.3 - 0.7 \log 0.7 \simeq 0.881 \text{ bit.}$$

$$n_2 = 2 \qquad q_{21} = 0.1, q_{22} = 0.2,$$

$$H_2 = -\frac{0.1}{0.3} \log \frac{0.1}{0.3} - \frac{0.2}{0.3} \log \frac{0.2}{0.3} \simeq 0.918 \text{ bit.}$$

$$H_3 \simeq 0.985 \text{ bit}, H_4 \simeq 0.918 \text{ bit}, H_5 = 0.$$

Verovatnosno n-arno stablo

Teorema 2.2 *Entropija listova*

Entropija listova verovatnosnog n-arnog stabla jednaka je sumi po svim unutrašnjim čvorovima uključujući koren, entropija grananja čvorova sa težinskim faktorima jednakim njihovim verovatnoćama, tj.

$$H_{leaf} = \sum_{i=1}^{M} P_i \, H_i$$

Primer 2.14

Ako nastavimo sa primerom 2.13, prema TEOREMI 2.2 važi

$$\begin{aligned} H_{leaf} &= 1 \cdot H_1 + 0.3 \cdot H_2 + 0.7 \cdot H_3 + 0.3 \cdot H_4 + 0.1 \cdot H_5 \\ &\simeq 0.881 + 0.3 \cdot 0.918 + 0.7 \cdot 0.985 + 0.3 \cdot 0.918 + 0 \text{ bit} \\ &\simeq 2.122 \text{ bit.} \end{aligned}$$

Što se slaže sa direktnim računanjem obavljenim u primeru 2.13.

Verovatnosno n-arno stablo

Teorema 2.3

Za bilo koja dva prefiksna koda nad istim informacionim izvorom, kod koji ima kraću očekivanu dužinu ima veću entropiju simbola.

Komprimovanje izvora informacija povećava entropiju simbola.