Kódelmélet

Témavázlat

- Gazdaságos kódolás
  - Kódolás
  - Betűnkénti kódolás, felbontható kód
    - Prefix kód
    - Blokk kód
    - Kódfa
  - A kódok hosszának alsó korlátja
    - McMillan-egyenlőtlenség
    - Kraft-tétele
  - Optimális kód
    - Átlagos szóhossz, a kód költsége
    - Optimális kódok
    - Entrópia
    - Shannon-tétele
    - Bináris Huffman-kód
    - Adaptív kódok: LZ-, LZW-kód

Gazdaságos kódolás

- A *hírközlésben* szükségünk van arra, hogy valamilyen *üzenetet* egy *csatornán* átjuttassunk. A csatorna azonban csak meghatározott jeleket tud befogadni, ezért az üzenetet időnként megfelelőképpen át kell alakítanunk, *kódolnunk* kell.
- Ez az átalakítás olyan kell legyen, hogy a csatorna túlsó oldalán többé-kevésbé helyesen *visszaállítható* legyen az eredeti üzenet.
- Az alábbiakban olyan kódolásokkal foglalkozunk, amelyek
  - lehetővé teszik a kódból az *üzenet helyes visszaállítását*, és
  - a kódok lehetőség szerint *rövidek*.
- Ennek az *elvi korlátait* vizsgáljuk.

Kódolás

Definíció.

Az $A=\{a_1,\dots, a_n\}$ véges, nemüres halmazt *ábécé*-nek nevezzük, elemei a *betűk*, a belőlük képezhető véges hosszú sorozatok a *szavak*. Az összes véges hosszú sorozat halmazát A^* jelöli.

Definíció.

Legyen B és C ábécé. A $f: B \to C^*$ leképezést *kódolásnak* nevezzük, ha injektív. $f(B)\subseteq C^*$ a kódszavak halmaza, a kód. A $b\in B$ *betű kódja* $f(b)$.

Az injektivitás garantálja a *dekódolhatóságot*, vagyis azt, hogy a képelemekből helyesen vissza tudjuk állítani a B halmaz elemeit.

1. példa.

Legyen $B=\{a, b, c\}$, $C=\{0, 1\}$ és $f(a)= 0$, $f(b)=01$, $f(c)=001$.

Ez a leképezés kódolás.

Betűnkénti kódolás, felbontható kód

Definíció.

Terjesszük ki f-et B*-ra a következőképpen:

Legyen $b=b_1b_2\ldots b_s \in B^*$. Ekkor $f(b)=f(b_1)f(b_2)\ldots f(b_s)$.
A B*-beli szavak kódját a szavakat alkotó betűk kódjainak egymás mellé írásával kapjuk. Ekkor az $f: B^* \to C^*$ kódolást *betűnkénti kódolásnak* nevezzük.

Definíció.

Az $f: B^* \to C^*$ betűnkénti kódolás *felbontható kódot* állít elő, ha két különböző B*-beli szóhoz tartozó kód különböző.

A kód felbonthatósága garantálja az üzenet kódjából az üzenet egyértelmű visszaállíthatóságát.

2. példa. Az 1. példa kódja például nem felbontható, mert

$f(ab)=f(c)$.

Prefix kód

Definíció.

Betűnkénti kódolás esetén a kódot *prefixnek* nevezzük, ha egyik kódszó sem valódi szókezdő része a másiknak.

3. példa. Az előző példában szereplő kód nem prefix.

f(a)= 0, f(b)=01, f(c)=001,

Például f(a) szerepel f(b) elején, más szóval f(b) az f(a)-nak folytatása.

4. példa.

Legyen B={a, b, c}, C={0, 1} és f(a)= 0, f(b)=10, f(c)=100.

Ez a kód prefix.

Tétel. Prefix kód felbontható.

Bizonyítás. Könnyen adható dekódolási algoritmus. Ez prefix kód esetén egyértelműen előállítja a kódolt üzenetből az eredetit.

Blokk-kód

Definíció.

Betűnkénti kódolás esetén a kódot *blokk-kódnak* nevezzük, ha a B halmaz mindegyik eleméhez ugyanolyan hosszú kódszó tartozik.

Tétel. Blokk-kód felbontható.

Bizonyítás. A blokk-kód egyúttal prefix kód is, így az előző tétel alkalmazható.

Kódfa

Definíció.

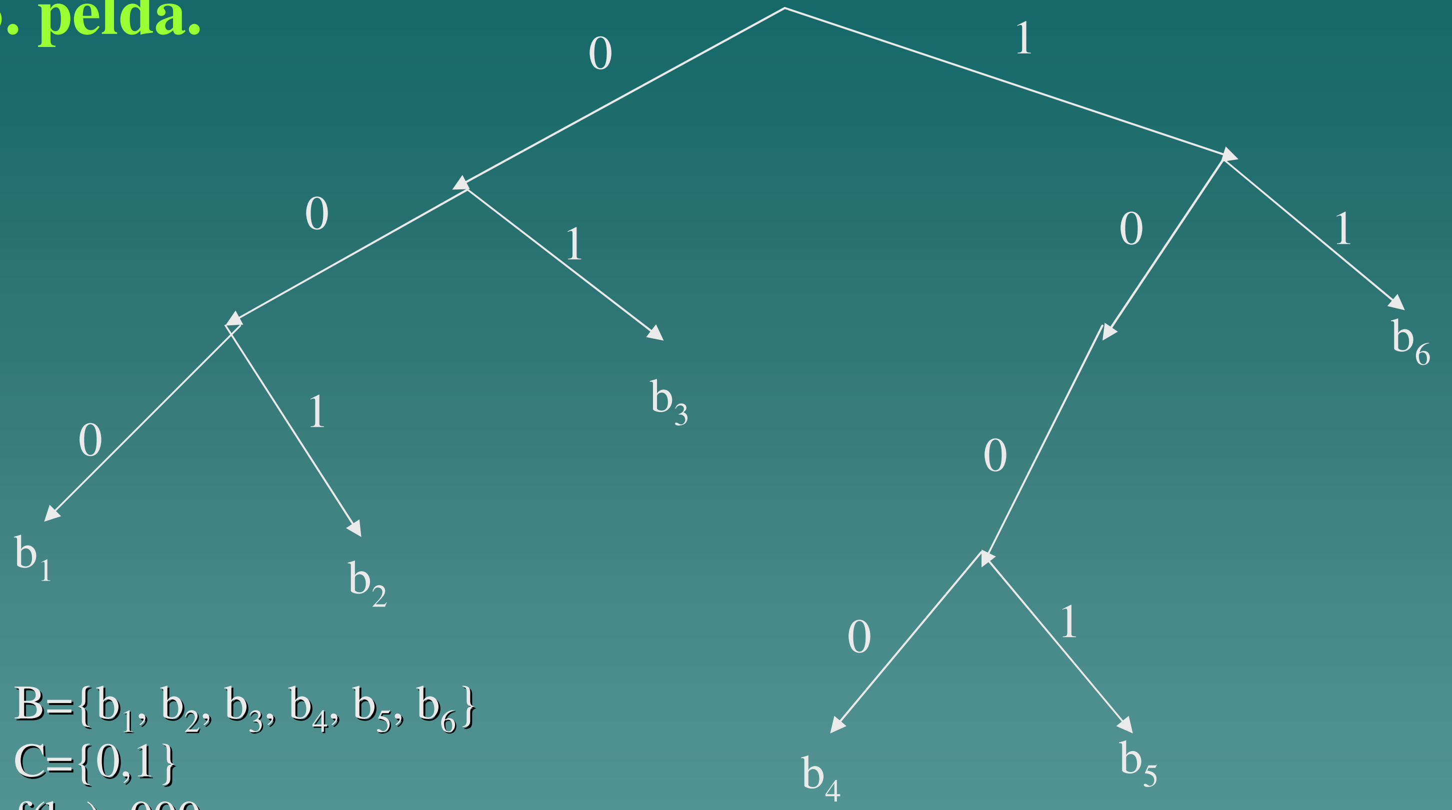
Az f: B → C* által létrehozott prefix kódhoz irányított fát, un. *kódfát* rendelhetünk a következő módon. A fa csúcsaiból legfeljebb |C| elemszámú él vezethet ki, ezeket az éleket C elemeivel címkézzük meg.

A fa leveleit (olyan csúcsok, amelyekből nem vezet ki él) a B elemeivel címkézzük meg. Ekkor a kódolás a következőképpen olvasható le a fáról.

Legyen az egyik levél a $b \in B$ és a gyökérből a hozzá vezető élek címkéi sorban $c_1, c_2, \ldots, c_k$. Ekkor $f(b)= c_1c_2\ldots c_k$.

A következő példában bináris prefix kód kódfáját látjuk.

5. példa.



$B=\{b_1, b_2, b_3, b_4, b_5, b_6\}$

$C=\{0,1\}$

$f(b_1)=000$

$f(b_2)=001$

$f(b_3)=01$

$f(b_4)=1000$

$f(b_5)=1001$

$f(b_6)=11$

A kódok hosszának alsó korlátja

McMillan-egyenlőtlenség

Tétel. (McMillan-egyenlőtlenség)

Tegyük fel, hogy f: B → C* felbontható kódot határoz meg.

B ={b_1, b_2, … , b_k}, és az f(b_1), f(b_2), …, f(b_k) kódszavak hossza {h_1, h_2, … , h_k}, |C|=c.

Ekkor $$\sum_{i=1}^{k} \frac{1}{c^{h_i}} \leq 1$$

6. példa. A 4. példa prefix kódjára |C|=2, a kódhosszak 1, 2, 3, a McMillan-egyenlőtlenség teljesül.

$$\sum_{i=1}^{k} \frac{1}{c^{h_i}} = \frac{1}{2} + \frac{1}{2^2} + \frac{1}{2^3} = \frac{7}{8} \leq 1$$

Vigyázzunk, a McMillan-egyenlőtlenség nem megfordítható. Ha teljesül az egyenlőtlenség, *nem biztos, hogy a kód felbontható*.

7. példa. Az 1. példa esetében ugyanazt az értéket kapjuk mint az 5. példában, pedig az előbbi kód nem felbontható.

Tétel. (Kraft-tétel)

Legyen B és C véges ábécé, B ={b_1, b_2, … , b_k}, |C|=c, és legyenek {h_1, h_2, … , h_k} pozitív egész számok, melyekre teljesül a McMillan-egyenlőtlenség.

$$\sum_{i=1}^{k} \frac{1}{c^{h_i}} \leq 1$$

Ekkor létezik olyan prefix kódot meghatározó f: B → C* kódolás, amelyre az f(b_1), f(b_2), …, f(b_k) kódszavak hossza éppen {h_1, h_2, … , h_k}.

Következmény.

A McMillan- és a Kraft-tételből következik, hogy ha $f: B \to C^*$ felbontható kódot határoz meg, akkor létezik olyan prefix kód, hogy a két kódban a B elemeihez tartozó kódszavak hossza megegyezik.

Ez a tény megnöveli a prefix kód jelentőségét.

Optimális kód

Átlagos szóhossz, a kód költsége

A kódolandó üzenetben a különböző jelek más és más gyakorisággal fordulhatnak elő. Ha törekszünk arra, hogy az üzenethez tartozó kód minél rövidebb legyen, akkor a gyakrabban előforduló jelekhez rövidebb kódot, míg a ritkábban előfordulókhoz a hosszabb kódokat érdemes rendelnünk.

Tegyük fel a továbbiakban, hogy az F jelforrás a $B = \{b_1, b_2, \dots, b_k\}$ ábécé jeleit egymástól függetlenül véletlenszerűen bocsátja ki. Jelölje p_i annak a valószínűségét, hogy az F által kibocsátott jel b_i. Feltesszük, hogy $p_i > 0$ ($i=1..k$), és

$$\sum_{i=1}^{k} p_i = 1$$

Elég hosszú, pl. M számú jelből álló jelsorozatban a benne előforduló b_i-k száma közelítőleg $p_i M$.

Az M számú jelből álló sorozat kódjának átlagos hossza: $M \sum_{i=1}^{k} p_i \cdot h_i$

Ha csökken a $\sum_{i=1}^{k} p_i \cdot h_i$ értéke, akkor csökken a közlések átlagos hossza is. Ez indokolja a következő definíciót.

Definíció.

Tegyük fel, hogy az f: B $\rightarrow$ C* felbontható kódolást alkalmazzuk, és az $f(b_1)$, $f(b_2)$, …, $f(b_k)$ kódszavak hossza $\{h_1, h_2, \ldots, h_k\}$, $|C|=c$, $K=f(B)\subseteq C^*$.
Jelölje p_i annak a valószínűségét, hogy az F forrás által kibocsátott jel b_i.

A *K kód F forrás melletti átlagos szóhossza*, vagy *költsége*:

$$H(K) = \sum_{i=1}^{k} p_i h_i$$

Optimális kódok

Definíció.

Legyen B és C véges ábécé. Rögzítsük a F jelforrást, vagyis a B ábécé betűihez tartozó p_i valószínűségeket. Tekintsük az $f: B \to C^*$ függvények által meghatározott felbontható kódokat. Ezek közül a legkisebb átlagos szóhosszúságú (költségű) kódot *optimális kódnak* nevezzük.

Korábbi megjegyzésünk alapján elég adott esetben az optimális prefix kódot keresnünk.

Entrópia

Definíció.

Az alábbi E(F) értéket az F forrás *entrópiájának* nevezzük. (A log 2-es alapú logaritmust jelöl)

$$E(F) = \sum_{i=1}^{k} p_i \log \frac{1}{p_i} = -\sum_{i=1}^{k} p_i \log p_i$$

$$H(K)=\sum_{i=1}^{k} p_i h_i \qquad E(F)=\sum_{i=1}^{k} p_i \log\frac{1}{p_i} = -\sum_{i=1}^{k} p_i \log p_i$$

Tétel. (Shannon tétele zajmentes csatornákra)

Egy F jelforráshoz tartozó tetszőleges K felbontható kódra teljesül a következő

$$H(K) \geq \frac{E(F)}{\log c}$$

Prefix kóddal ez a korlát jól megközelíthető.

Tétel.

Létezik olyan f: B→C* prefix kód, amelyre $H(K) \leq \frac{E(F)}{\log c} + 1$

Bizonyítás.

A bizonyítást például az un. Shannon-Fano kód segítségével lehet elvégezni.

Bináris Huffman-kód

Legyen $B = \{b_1, b_2, \dots, b_k\}$, a valószínűségek pedig sorban (nagyság szerint csökkenően rendezve):

$\{0,20;\ 0,20;\ 0,19;\ 0,12;\ 0,11;\ 0,09;\ 0,09\}$

b_1	0,20	0,20	0,23	0,37	0,40	0,60
b_2	0,20	0,20	0,20	0,23	0,37	0,40
b_3	0,19	0,19	0,20	0,20	0,23	
b_4	0,12	0,18	0,19	0,20		
b_5	0,11	0,12	0,18			
b_6	0,09	0,11				
b_7	0,09					

b_1	10	10	01	00	1	0
b_2	11	11	10	01	00	1
b_3	000	000	11	10	01	
b_4	010	001	000	11		
b_5	011	010	001			
b_6	0010	011				
b_7	0011					

A Huffmann-kód előnye:

Optimális kódot állít elő.

A Huffmann-kód hátrányai:

- Ismernünk kell kódolásnál a teljes szöveget.
- Kétszer kell végigmennünk az adatokon. Először meghatározzuk a forrásbetűk relatív gyakoriságát, ami megegyezik a valószínűségekkel, majd ennek felhasználásával elvégezzük a tényleges kódolást.

Adaptív Huffmann-kódolás.

Csak egyszer megy végig az adatokon. Az optimalitás rovására időt takarítunk meg. Egy forrásbetűt az előző forrásbetűk előfordulásai alapján kódolunk, s ezzel együtt lépésenként változik maga a kód is. Az aktuális forrásbetű kódolását egy, az előzőleg feldolgozott forrásbetűkre optimális kóddal hajtjuk végre.

Adaptív kódok.

Menet közben gyűjtünk információt a forrásszimbólumokról, az aktuális szimbólumot az ezt megelőző szimbólumok alapján kódoljuk.

Lempel-Ziv kódok:

- **LZ77 algoritmus.** 1977-ben publikálták.
- **LZ78 algoritmus.**

LZW kód:

Terry Welch az LZ78-at módosította.

Az Unix COMPRESS parancsa és a GIF (Graphics Interchange Format) képtömörítő is az LZW algoritmust használja.

Irodalomjegyzék

- Demetrovics, Denev, Pavlov: *A számítástudomány matematikai alapjai* Tankönyvkiadó, Budapest, 1985
- Győrfi László-Győri Sándor-Vajda István: *Információ és kódelmélet* Typotex Kiadó, 2000
- Jablonszkij, Lupanov: *Diszkrét matematika a számítástudományban* Műszaki Könyvkiadó, 1980