Információelmélet

Informatika elméleti alapjai

Horváth Árpád

Óbudai Egyetem
Alba Regia Műszaki Kar (AMK)
Székesfehérvár

2014. október 26.

Vázlat

Információelmélet

Egy jelsorozat esetén vizsgáljuk, mennyi információt tartalmaz.
Nem érdekel minket a jelek tényleges jelentése.
Ahonnan a jelek jönnek, forrásnak nevezzük. A jelek összességét a forrás jelkészletének (szimbólumkészletének, ábécéjének) nevezzük.
Egyszerűség kedvéért feltételezzük, hogy véges számú jel van (diszkrét jelkészlet). Ezt nevezzük diszkrét forrásnak.

Vázlat

Mekkora egy jel információtartalma?

Mennyi igen-nem válasszal határozhatunk meg egy kártyát a magyarkártya-pakliból?
Makk VIII-as (M,VIII)

M,VII	M,VIII	M,IX	M,X	M,A	M,F	M,K	M,Á
T,VII	T,VIII	T,IX	T,X	T,A	T,F	T,K	T,Á
Z,VII	Z,VIII	Z,IX	Z,X	Z,A	Z,F	Z,K	Z,Á
P,VII	P,VIII	P,IX	P,X	P,A	P,F	P,K	P,Á

M,VII	M,VIII	M,IX	M,X	M,A	M,F	M,K	M,Á
T,VII	T,VIII	T,IX	T,X	T,A	T,F	T,K	T,Á
Z,VII	Z,VIII	Z,IX	Z,X	Z,A	Z,F	Z,K	Z,Á
P,VII	P,VIII	P,IX	P,X	P,A	P,F	P,K	P,Á

M,VII	M,VIII	M,IX	M,X	M,A	M,F	M,K	M,Á
T,VII	T,VIII	T,IX	T,X	T,A	T,F	T,K	T,Á
Z,VII	Z,VIII	Z,IX	Z,X	Z,A	Z,F	Z,K	Z,Á
P,VII	P,VIII	P,IX	P,X	P,A	P,F	P,K	P,Á

M,VII	M,VIII	M,IX	M,X	M,A	M,F	M,K	M,Á
T,VII	T,VIII	T,IX	T,X	T,A	T,F	T,K	T,Á
Z,VII	Z,VIII	Z,IX	Z,X	Z,A	Z,F	Z,K	Z,Á
P,VII	P,VIII	P,IX	P,X	P,A	P,F	P,K	P,Á

M,VII	M,VIII	M,IX	M,X	M,A	M,F	M,K	M,Á
T,VII	T,VIII	T,IX	T,X	T,A	T,F	T,K	T,Á
Z,VII	Z,VIII	Z,IX	Z,X	Z,A	Z,F	Z,K	Z,Á
P,VII	P,VIII	P,IX	P,X	P,A	P,F	P,K	P,Á

M,VII	M,VIII	M,IX	M,X	M,A	M,F	M,K	M,Á
T,VII	T,VIII	T,IX	T,X	T,A	T,F	T,K	T,Á
Z,VII	Z,VIII	Z,IX	Z,X	Z,A	Z,F	Z,K	Z,Á
P,VII	P,VIII	P,IX	P,X	P,A	P,F	P,K	P,Á

Kétféle modellt vizsgálunk a források esetén

Legyen egy diszkrét forrás esetén egy jel előfordulási valószínűsége független az előző jel(ek) értékétől. (független=independent)

- IID modell: A jelek egyforma valószínűséggel jelennek meg, tehát ha 32 jelem van, mindegyik 1/32 valószínűséggel.

Kétféle modellt vizsgálunk a források esetén

Legyen egy diszkrét forrás esetén egy jel előfordulási valószínűsége független az előző jel(ek) értékétől. (független=independent)

- IID modell: A jelek egyforma valószínűséggel jelennek meg, tehát ha 32 jelem van, mindegyik 1/32 valószínűséggel. (independent, identically distributed).

Kétféle modellt vizsgálunk a források esetén

Legyen egy diszkrét forrás esetén egy jel előfordulási valószínűsége független az előző jel(ek) értékétől. (független=independent)

- IID modell: A jelek egyforma valószínűséggel jelennek meg, tehát ha 32 jelem van, mindegyik 1/32 valószínűséggel. (independent, identically distributed).
- IRD modell: A jelek tetszőleges valószínűséggel jelennek meg, tehát hiába van 32 jelem (kártyám), lehet az egyik (pl. piros király) valószínűsége 1/2 is.
  Fontos, hogy ebben az esetben minden pillanatban 1/2 lesz a valószínűsége az adott jelnek, akármilyen jelek is voltak előtte.)

Kétféle modellt vizsgálunk a források esetén

Legyen egy diszkrét forrás esetén egy jel előfordulási valószínűsége független az előző jel(ek) értékétől. (független=independent)

- IID modell: A jelek egyforma valószínűséggel jelennek meg, tehát ha 32 jelem van, mindegyik 1/32 valószínűséggel. (independent, identically distributed).
- IRD modell: A jelek tetszőleges valószínűséggel jelennek meg, tehát hiába van 32 jelem (kártyám), lehet az egyik (pl. piros király) valószínűsége 1/2 is.
  (Fontos, hogy ebben az esetben minden pillanatban 1/2 lesz a valószínűsége az adott jelnek, akármilyen jelek is voltak előtte.)

Kétféle modellt vizsgálunk a források esetén

Legyen egy diszkrét forrás esetén egy jel előfordulási valószínűsége független az előző jel(ek) értékétől. (független=independent)

- IID modell: A jelek egyforma valószínűséggel jelennek meg, tehát ha 32 jelem van, mindegyik 1/32 valószínűséggel. (independent, identically distributed).
- IRD modell: A jelek tetszőleges valószínűséggel jelennek meg, tehát hiába van 32 jelem (kártyám), lehet az egyik (pl. piros király) valószínűsége 1/2 is.
  (Fontos, hogy ebben az esetben minden pillanatban 1/2 lesz a valószínűsége az adott jelnek, akármilyen jelek is voltak előtte.)
  (independent, randomly distributed).

Random jelentései

- véletlen

Lásd még RAM: random access memory. (Nem LIFO.)

Random jelentései

- véletlen
- tetszőleges

Lásd még RAM: random access memory. (Nem LIFO.)

Random jelentései

- véletlen
- tetszőleges
- rendszertelen

Lásd még RAM: random access memory. (Nem LIFO.)

Random jelentései

- véletlen
- tetszőleges
- rendszertelen
- találomra történő

Lásd még RAM: random access memory. (Nem LIFO.)

Random jelentései

- véletlen
- tetszőleges
- rendszertelen
- találomra történő

Lásd még RAM: random access memory. (Nem LIFO.)

IID

A 32 kártyából, ha mindegyiket egyforma valószínűséggel választhatják ki, akkor 5 igen-nem válaszból kitalálhatom a kártyát. Általában igaz, ha 2^m lehetőségem van, akkor m kérdést szükséges feltennem.
$X = \{x_1, x_2, \ldots, x_n\}$ jelhalmaz esetén $\log_2 n$ kérdés kell (felfele kerekítve a következő egészre).
Egy jel (pl. véletlen választott kártya) információtartalma az IID modellben

$$I = \log_2 n$$

IRD

Ha nem egyforma valószínűsége van egy-egy kártyának (jelnek), akkor lehet, hogy érdemesebb a kártyaszám-felezéses taktika helyett másikat választani. Inkább a valószínűségeket kell felezni, mint a kártyák (jelek) számát.
Ha pl. a Piros VIII-as jelenik meg az esetek felében, akkor érdemes lehet arra rákérdezni, így azt egy kérdéssel kitaláljuk. Ha így teszek, akkor lesznek olyan kártyalapok, amelyhez 5-nél több kérdés kell majd, de ha ezek elég ritkán jelennek meg, akkor az átlagos kérdésszám lehet öt alatti.
A ritkábban megjelenő kártya megjelenésének az információtartalma nagyobb. Az, hogy 4 lapom közül mind a négy ász, az nagyobb információtartalmú, mint hogy minden kártyám számos (VII-es, VIII-as, IX-es vagy X-es),

IRD

Ha nem egyforma valószínűsége van egy-egy jelnek, akkor a kisebb valószínűségűnek nagyobb az információtartalma.
$X = \{x_1, x_2, \ldots, x_n\}$ jelekhez tartozzanak $P(X) = \{p_1, p_2, \ldots, p_n\}$ valószínűségek.
Legyen a rendszer teljes:

$$p_1 + p_2 + \ldots + p_n = \sum_{k=1}^{n} p_k = 1 = 100\%$$

Ekkor minden x_k értékhez valamilyen p_k-tól függő információtartalom tartozik.
Az x_k jel információtartalma az IRD modellben:

$$I_k = \log_2 \frac{1}{p_k} = -\log_2 p_k$$

(Monotonitás, IID.)

Az órán vizsgáltuk a következő eseteket

$$X = \{A, B, C, D\}$$

jelhalmaz esetén.

- $P(X) = \{\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\}$

Az órán vizsgáltuk a következő eseteket

$$X = \{A, B, C, D\}$$

jelhalmaz esetén.

- $P(X) = \{\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\}$
- $P(X) = \{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\}$

Az órán vizsgáltuk a következő eseteket

$$X = \{A, B, C, D\}$$

jelhalmaz esetén.

- $P(X) = \{\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\}$
- $P(X) = \{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\}$
- $P(X) = \{0.8, 0.1, 0.05, 0.05\}$

Az órán vizsgáltuk a következő eseteket

$$X = \{A, B, C, D\}$$

jelhalmaz esetén.

- $P(X) = \{\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\}$
- $P(X) = \{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\}$
- $P(X) = \{0.8, 0.1, 0.05, 0.05\}$

Egy-egy kódot adtam meg az egyes betűkhöz, amivel vizsgáltuk a jelenként felhasználandó bitek számának várható értékét.
Az első esetben az alábbi első eloszlás, a második kettőben az alábbi második kódokkal vizsgáltuk.

$$A = 00, B = 01, C = 10, D = 11$$

$$A = 1, B = 01, C = 001, D = 000$$

Az információ egységei

Az eddigiekben a kettes alapszám helyett mást is használhatunk:

alapszám (a)	egység
2	bit (Shannon)
10	Hartley
e	Nat

$$I_k = \log_a \frac{1}{p_k} = -\log_a p_k$$

2-es és más alapú logaritmus kiszámítása számológéppel:

$$\log_2 x = \frac{\lg x}{\lg 2} \qquad \log_a x = \frac{\lg x}{\lg a}$$

$$p_k = 0{,}1 \Rightarrow I_k = \lg \frac{1}{0{,}1} = \lg 10 = 1 \text{ Hartley.}$$

$$p_k = 0{,}1 \Rightarrow I_k = \log_2 10 = \frac{\lg 10}{\lg 2} = 3{,}3219 \text{ bit.}$$

Vázlat

Entrópia

Mekkora a várható értéke a következő jel információtartalmának?
A valószínűségekkel súlyozni kell az összes jel információtartalmát.
Ezt nevezzük entrópiának:

$$H(X) = \sum_{k=1}^{n} p_k \cdot I_k = \sum_{k=1}^{n} p_k \cdot \log_2 \frac{1}{p_k} \quad \text{bit/jel}$$

IID modellnél:

$$H_{n,IID} = n \cdot \frac{1}{n} \cdot \log_2 n = \log_2 n \quad \text{bit/jel}$$

Mire jó ez nekünk? Meghatározhatjuk, mennyi bit szükséges feltétlenül az információ átviteléhez.

Példa

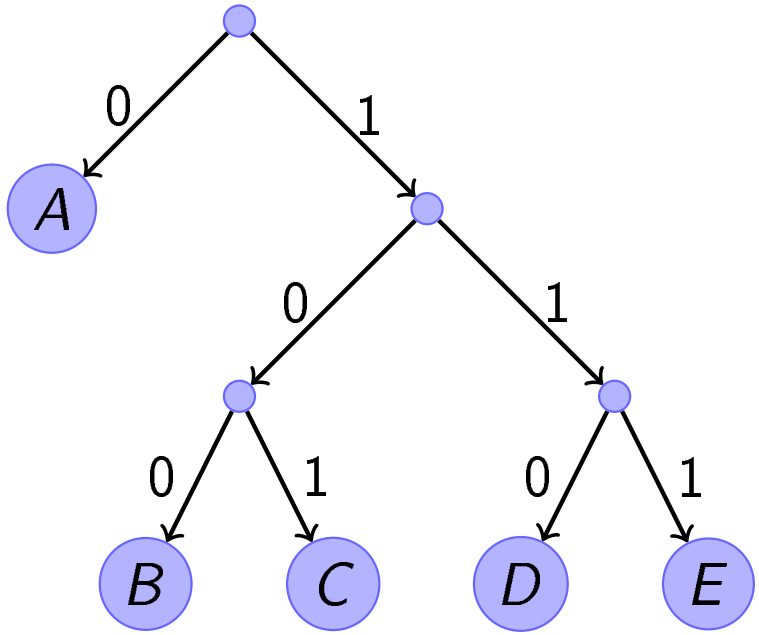
- $$X = \{A, B, C, D, E\}, \quad P(X) = \left\{\frac{1}{2}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}\right\}$$

Példa

- $$X = \{A, B, C, D, E\}, \quad P(X) = \left\{\frac{1}{2}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}\right\}$$
- $$I_A = 1bit, \quad I_B = I_C = I_D = I_E = 3bit,$$
$$H(X) = \left(\frac{1}{2} \cdot 1 + 4 \cdot \frac{1}{8} \cdot 3\right) \quad \text{bit/jel} = 2 \quad \text{bit/jel}$$

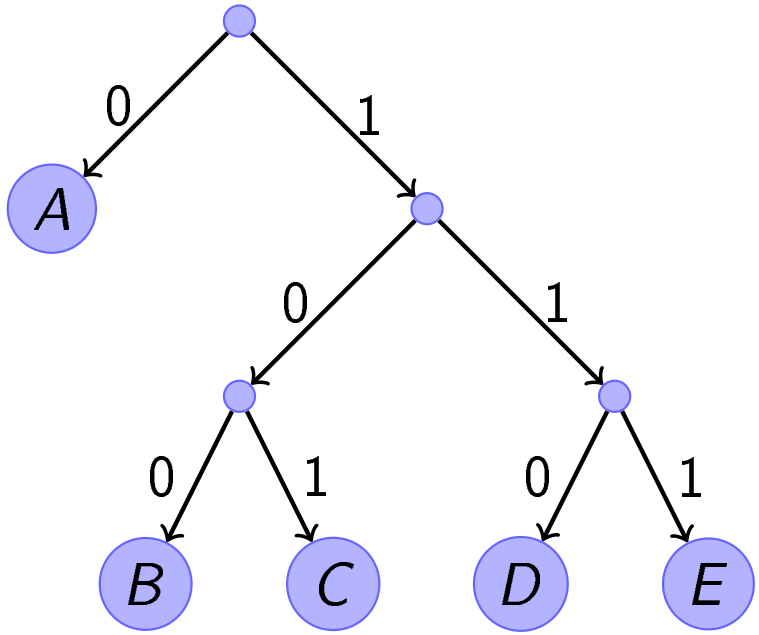
Példa

- Legyen A=0, B=100, C=101, D=110, E=111.



Példa

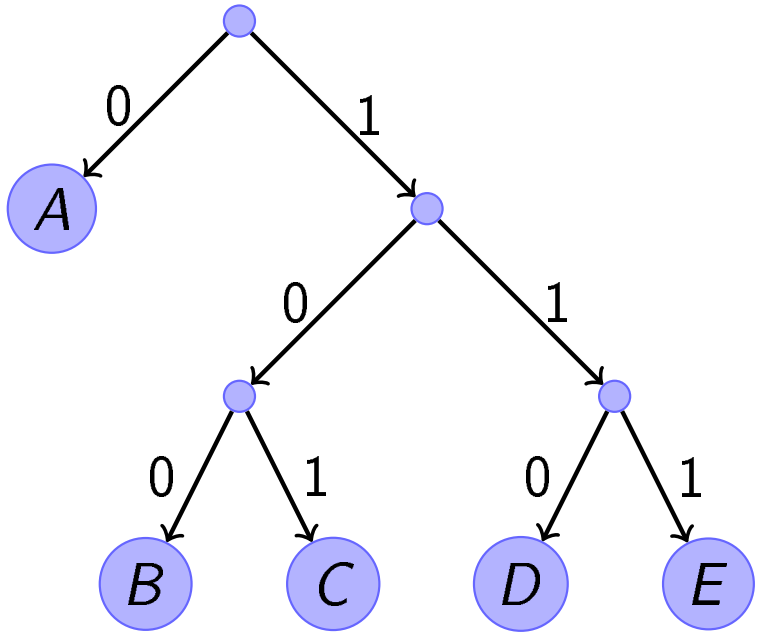
- Legyen A=0, B=100, C=101, D=110, E=111.



- Minden bitsorozat egyértelműen visszafejthető pl: 1000001101110101

Példa

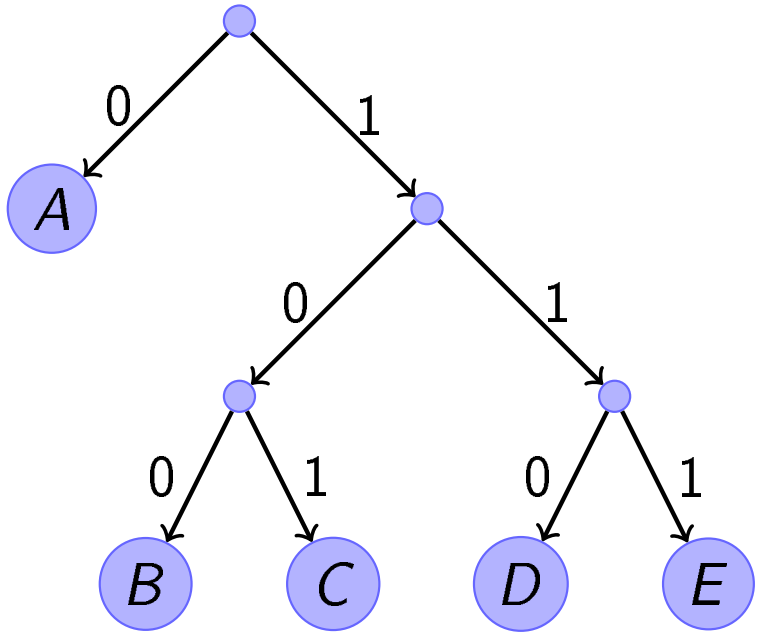
- Legyen A=0, B=100, C=101, D=110, E=111.



- Minden bitsorozat egyértelműen visszafejthető pl: 1000001101110101
- 100,0,0,0,110,111,0,101 = BAAADEAC

Példa

- Legyen A=0, B=100, C=101, D=110, E=111.



- Minden bitsorozat egyértelműen visszafejthető pl: 1000001101110101
- 100,0,0,0,110,111,0,101 = BAAADEAC
- Ekkor 8 karakterből átlagosan 4 db „A” (4 bit) a többi négy 3 bites (12 bit): ez összesen 16 bit $\Rightarrow$ 2 bit/jel.

További mérőszámok

- Mindig az IID esetben a legnagyobb az entrópia azonos jelszám (n) esetén:

$$H_{n,max} = H_{n,IID} \geq H(X), \qquad |X| = n$$

További mérőszámok

- Mindig az IID esetben a legnagyobb az entrópia azonos jelszám (n) esetén:

$$H_{n,max} = H_{n,IID} \geq H(X), \qquad |X| = n$$

- Hatásfok: Az entrópia aránya az ugyanannyi jelet tartalmazó IID modelléhez viszonyítva.

$$\eta = \frac{H(X)}{H_{n,max}}$$

További mérőszámok

- Mindig az IID esetben a legnagyobb az entrópia azonos jelszám (n) esetén:

$$H_{n,max} = H_{n,IID} \geq H(X), \qquad |X| = n$$

- Hatásfok: Az entrópia aránya az ugyanannyi jelet tartalmazó IID modelléhez viszonyítva.

$$\eta = \frac{H(X)}{H_{n,max}}$$

- Redundancia: (hétköznapi jelentés: terjengősség)

$$R = 1 - \eta$$

Példa

- $$X = \{A, B, C, D, E\}, \quad P(X) = \left\{\frac{1}{2}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}\right\}$$

Példa

- $$X = \{A, B, C, D, E\}, \quad P(X) = \left\{\frac{1}{2}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}\right\}$$
- Hatásfok:

$$\eta = \frac{H(X)}{H_{n,max}} = \frac{2 \quad \text{bit/jel}}{\log_2 5 \quad \text{bit/jel}} = \frac{2 \quad \text{bit/jel}}{2,3219 \quad \text{bit/jel}} = 0,8614 \approx 86\%$$

Példa

- $$X = \{A, B, C, D, E\}, \quad P(X) = \left\{\frac{1}{2}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}\right\}$$
- Hatásfok:

$$\eta = \frac{H(X)}{H_{n,max}} = \frac{2 \quad \text{bit/jel}}{\log_2 5 \quad \text{bit/jel}} = \frac{2 \quad \text{bit/jel}}{2,3219 \quad \text{bit/jel}} = 0,8614 \approx 86\%$$

- Redundancia:

$$R = 1 - \eta = 0,1386 \approx 14\%$$

Ajánlott segédlet

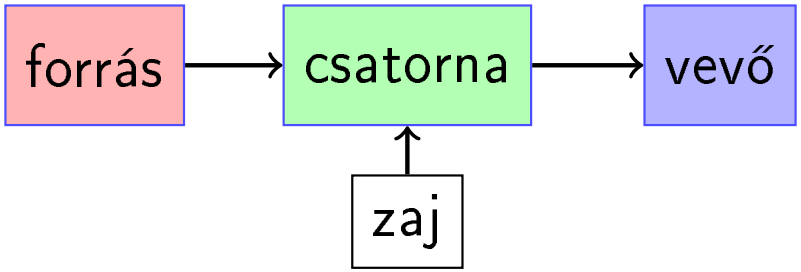
Dr. Tóth Mihály – Tóth Gergely: Az információ– és kódoláselmélet
Kidolgozott példák és feladatok: 1.13.1, 1.13.2, 1.14.28, 1.14.35, 1.14.36, 1.14.37, 1.14.38, 1.14.39,
A feladatok a jegyzet I. részének 24. oldalán kezdődnek.

Első rész vége

Köszönöm a figyelmet!

Vázlat

Információátvitel diagrammja



Példák:

Beszélgetés: száj ⇒ levegő ⇒ fül

Földfelszíni rádióadás:
a rádióadó vagy az átjátszó antennája ⇒ az „éter” ⇒ a rádió antennája

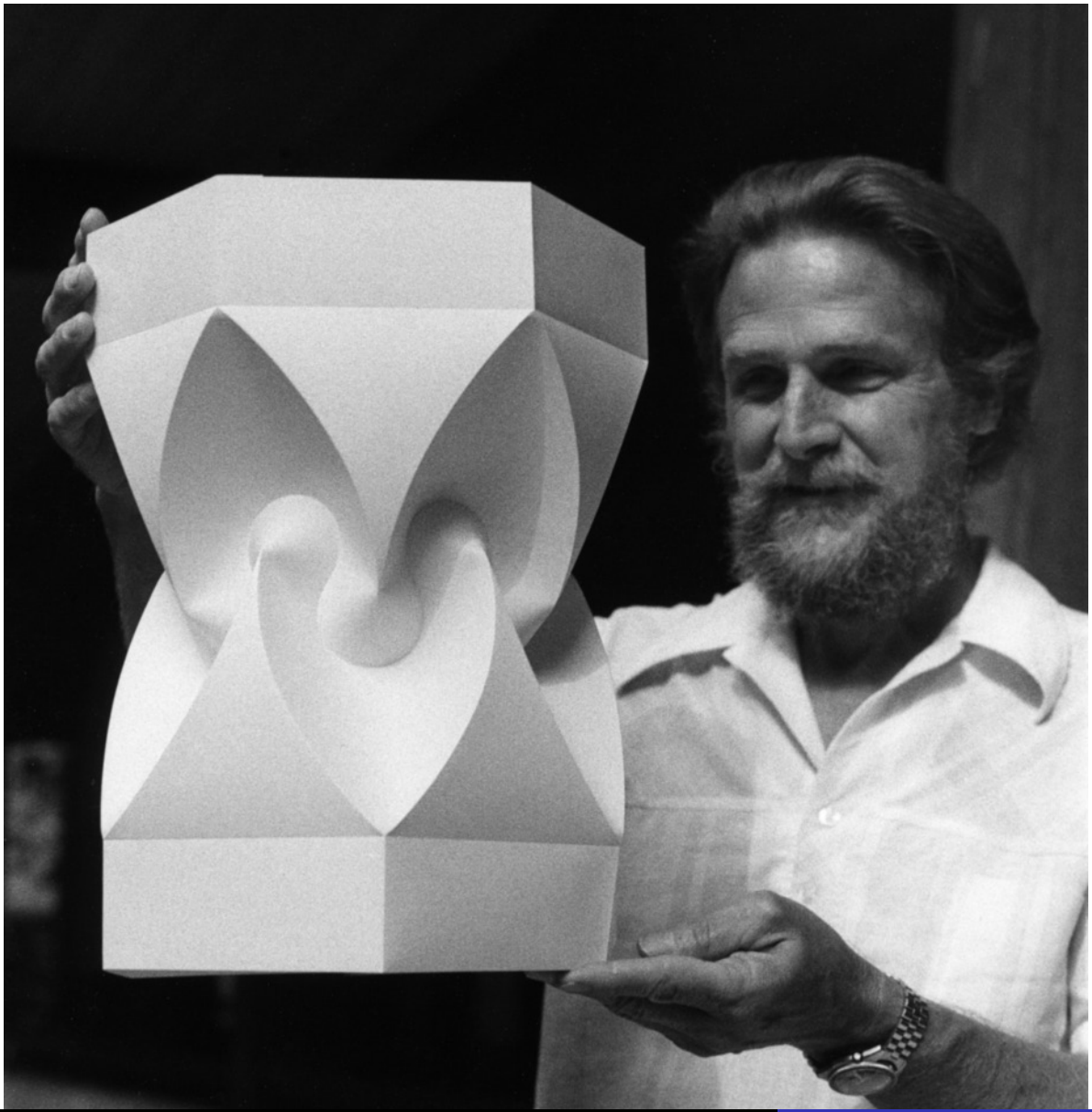
Internet: router1 ⇒ üvegszál ⇒ router2

koncert ⇒ CD ⇒ emberi fül

Forráskódolás: a forrás által küldött jelek kódolása legkisebb redundanciával („legkevesebb biten”)
Csatornakódolás: A jelet úgy kódolom, hogy a hibákat észre tudjam venni vagy javítani tudjam.

Vázlat

Huffman-kódolás (1952), David Huffman (1925-1999)



Feladat

Kódoljuk az alábbi eloszlású IRD forrást Huffman-kódolással!

k karakter	n_k darabszám
A	45
B	13
C	12
D	16
E	9
F	5

A Huffman-kódolás algoritmusa

1. Növekvő gyakoriságú sorrendbe rakom a jeleket.
   Azonos gyakoriságnál ABC-rendbe (vagy pl. ASCII-kódbeli rendbe).
   Kezdetben 1 jel 1 fának számít.

A Huffman-kódolás algoritmusa

1. Növekvő gyakoriságú sorrendbe rakom a jeleket.
   Azonos gyakoriságnál ABC-rendbe (vagy pl. ASCII-kódbeli rendbe).
   Kezdetben 1 jel 1 fának számít.
2. Ciklus, amíg van legalább két fa

Ciklus vége

A Huffman-kódolás algoritmusa

1. Növekvő gyakoriságú sorrendbe rakom a jeleket.
   Azonos gyakoriságnál ABC-rendbe (vagy pl. ASCII-kódbeli rendbe).
   Kezdetben 1 jel 1 fának számít.
2. Ciklus, amíg van legalább két fa
   1. Az első kettő fát összekötöm egy új gyökérrel.
      Gyakorisága a két fa gyakoriságának összege lesz.
      A baloldalt 0-val, a jobbot 1-gyel cimkézem.

   Ciklus vége

A Huffman-kódolás algoritmusa

1. Növekvő gyakoriságú sorrendbe rakom a jeleket.
   Azonos gyakoriságnál ABC-rendbe (vagy pl. ASCII-kódbeli rendbe).
   Kezdetben 1 jel 1 fának számít.
2. Ciklus, amíg van legalább két fa
   1. Az első kettő fát összekötöm egy új gyökérrel.
      Gyakorisága a két fa gyakoriságának összege lesz.
      A baloldalt 0-val, a jobbot 1-gyel cimkézem.
   2. A következő ábrán növekvő gyakorisági sorrendbe rakom a fákat. (Felül hagyok helyet az összekötésnek.)
      Azonos értékeknél az újat lehető leghátulra rakom.

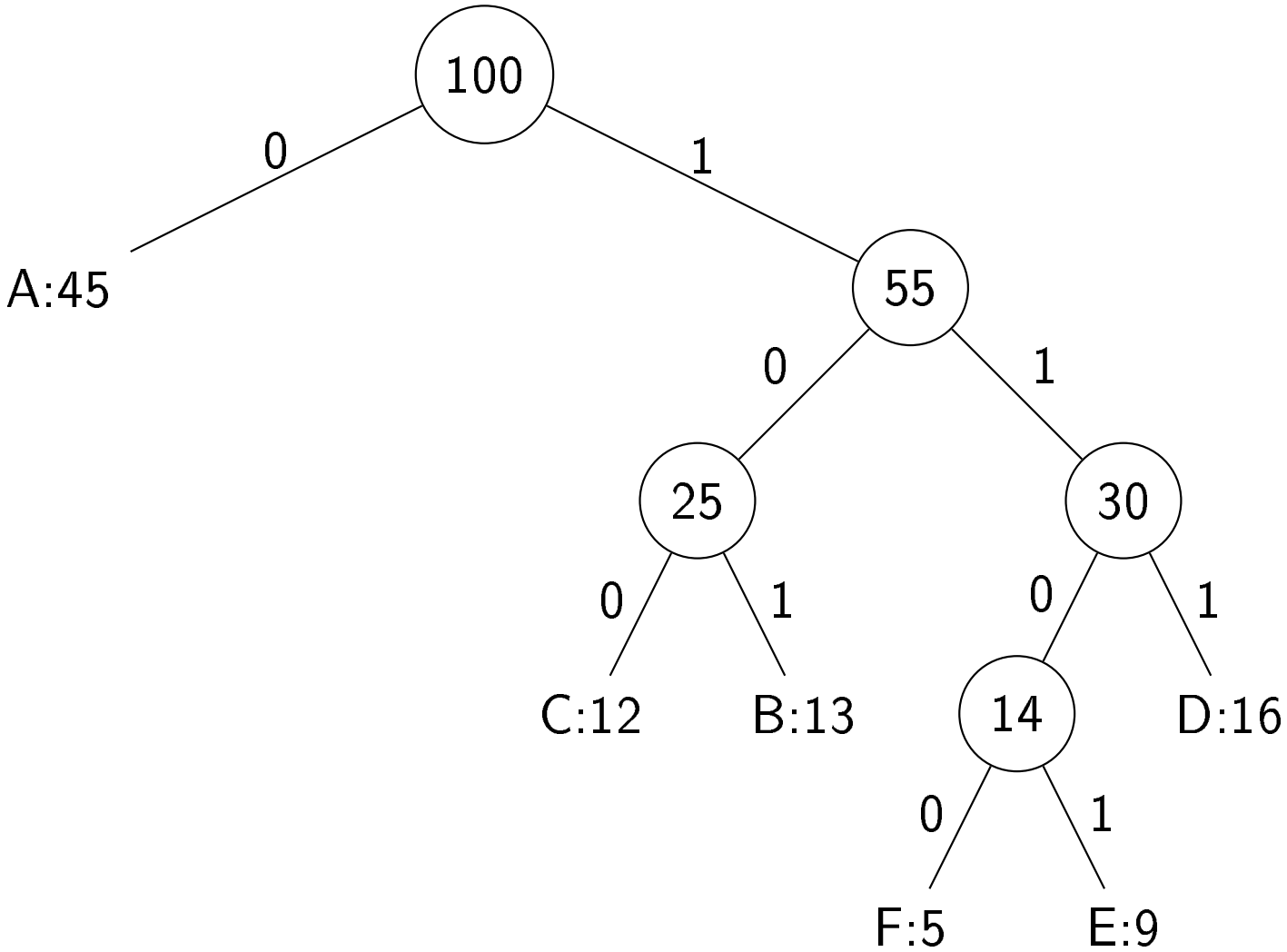
   Ciklus vége

A Huffman-kódolás algoritmusa

1. Növekvő gyakoriságú sorrendbe rakom a jeleket.
   Azonos gyakoriságnál ABC-rendbe (vagy pl. ASCII-kódbeli rendbe).
   Kezdetben 1 jel 1 fának számít.
2. Ciklus, amíg van legalább két fa
   1. Az első kettő fát összekötöm egy új gyökérrel.
      Gyakorisága a két fa gyakoriságának összege lesz.
      A baloldalt 0-val, a jobbot 1-gyel cimkézem.
   2. A következő ábrán növekvő gyakorisági sorrendbe rakom a fákat. (Felül hagyok helyet az összekötésnek.)
      Azonos értékeknél az újat lehető leghátulra rakom.

   Ciklus vége
3. Leolvasom a jelek kódjait.

Huffman-kód fája



B kódja: 101, F kódja: 1100

ASCII-tábla, 7 bites

	0	1	2	3	4	5	6	7
0	NUL	DLE	SP	0	@	P	‘	p
1	SOH	DC1	!	1	A	Q	a	q
2	STX	DC2	"	2	B	R	b	r
3	ETX	DC3	#	3	C	S	c	s
4	EOT	DC4	$	4	D	T	d	t
5	ENQ	NAK	%	5	E	U	e	u
6	ACK	SYN	&	6	F	V	f	v
7	BEL	ETB	’	7	G	W	g	w
8	BS	CAN	(	8	H	X	h	x
9	HT	EM	)	9	I	Y	i	y
A	LF	SUB	*	:	J	Z	j	z
B	VT	ESC	+	;	K	[	k	{
C	FF	FS	,	<	L	\	l	|
D	CR	GS	–	=	M	]	m	}
E	SO	RS	.	>	N	^	n	~
F	SI	US	/	?	O	__	o	DEL

Q:
9:
: 0x20
:
0x57414C4C

Szóköz (space=SP), számok, nagybetűk, kisbetűk helye

ASCII-tábla, 7 bites

	0	1	2	3	4	5	6	7
0	NUL	DLE	SP	0	@	P	‘	p
1	SOH	DC1	!	1	A	Q	a	q
2	STX	DC2	"	2	B	R	b	r
3	ETX	DC3	#	3	C	S	c	s
4	EOT	DC4	$	4	D	T	d	t
5	ENQ	NAK	%	5	E	U	e	u
6	ACK	SYN	&	6	F	V	f	v
7	BEL	ETB	’	7	G	W	g	w
8	BS	CAN	(	8	H	X	h	x
9	HT	EM	)	9	I	Y	i	y
A	LF	SUB	*	:	J	Z	j	z
B	VT	ESC	+	;	K	[	k	{
C	FF	FS	,	<	L	\	l	|
D	CR	GS	–	=	M	]	m	}
E	SO	RS	.	>	N	^	n	~
F	SI	US	/	?	O	__	o	DEL

Q: 0x51
9:
: 0x20
:
0x57414C4C

Szóköz (space=SP), számok, nagybetűk, kisbetűk helye

ASCII-tábla, 7 bites

	0	1	2	3	4	5	6	7
0	NUL	DLE	SP	0	@	P	‘	p
1	SOH	DC1	!	1	A	Q	a	q
2	STX	DC2	"	2	B	R	b	r
3	ETX	DC3	#	3	C	S	c	s
4	EOT	DC4	$	4	D	T	d	t
5	ENQ	NAK	%	5	E	U	e	u
6	ACK	SYN	&	6	F	V	f	v
7	BEL	ETB	’	7	G	W	g	w
8	BS	CAN	(	8	H	X	h	x
9	HT	EM	)	9	I	Y	i	y
A	LF	SUB	*	:	J	Z	j	z
B	VT	ESC	+	;	K	[	k	{
C	FF	FS	,	<	L	\	l	|
D	CR	GS	–	=	M	]	m	}
E	SO	RS	.	>	N	^	n	~
F	SI	US	/	?	O	__	o	DEL

Q: 0x51
9: 0x39
: 0x20
:
0x57414C4C

Szóköz (space=SP), számok, nagybetűk, kisbetűk helye

ASCII-tábla, 7 bites

	0	1	2	3	4	5	6	7
0	NUL	DLE	SP	0	@	P	‘	p
1	SOH	DC1	!	1	A	Q	a	q
2	STX	DC2	"	2	B	R	b	r
3	ETX	DC3	#	3	C	S	c	s
4	EOT	DC4	$	4	D	T	d	t
5	ENQ	NAK	%	5	E	U	e	u
6	ACK	SYN	&	6	F	V	f	v
7	BEL	ETB	’	7	G	W	g	w
8	BS	CAN	(	8	H	X	h	x
9	HT	EM	)	9	I	Y	i	y
A	LF	SUB	*	:	J	Z	j	z
B	VT	ESC	+	;	K	[	k	{
C	FF	FS	,	<	L	\	l	|
D	CR	GS	–	=	M	]	m	}
E	SO	RS	.	>	N	^	n	~
F	SI	US	/	?	O	__	o	DEL

Q: 0x51
9: 0x39
szóköz: 0x20
:
0x57414C4C

Szóköz (space=SP), számok, nagybetűk, kisbetűk helye

ASCII-tábla, 7 bites

	0	1	2	3	4	5	6	7
0	NUL	DLE	SP	0	@	P	‘	p
1	SOH	DC1	!	1	A	Q	a	q
2	STX	DC2	"	2	B	R	b	r
3	ETX	DC3	#	3	C	S	c	s
4	EOT	DC4	$	4	D	T	d	t
5	ENQ	NAK	%	5	E	U	e	u
6	ACK	SYN	&	6	F	V	f	v
7	BEL	ETB	’	7	G	W	g	w
8	BS	CAN	(	8	H	X	h	x
9	HT	EM	)	9	I	Y	i	y
A	LF	SUB	*	:	J	Z	j	z
B	VT	ESC	+	;	K	[	k	{
C	FF	FS	,	<	L	\	l	|
D	CR	GS	–	=	M	]	m	}
E	SO	RS	.	>	N	^	n	~
F	SI	US	/	?	O	__	o	DEL

Q: 0x51
9: 0x39
szóköz: 0x20
WALL: 0x57414C4C

Szóköz (space=SP), számok, nagybetűk, kisbetűk helye

Ha a következő karaktersorozatot a Huffman-algoritmussal tömörítem, akkor mi lesz az egyes jelekhez rendelt bitsorozat?

BENEDEK␣ELEK

Az üzenethez tartozó bitsorozat:

A Huffman-kódban hány bit jut átlagosan egy jelre? ($\langle d_k \rangle$)
Mekkora lesz az egyes karakterek egyedi információtartalma (I_k)?
Milyen alsó határt számolhatunk a szükséges bitek számára (H)?
Mi lesz a 10101001110110011110 bitsorozathoz tartozó üzenet?

A Huffman-kóddal kapcsolatos eredmények

k	n_k	kód	d_k kódhossz [bit]	$n_k d_k$
␣	1	1100	4	4
B	1	1101	4	4
D	1	1110	4	4
L	1	1111	4	4
N	1	100	3	3
K	2	101	3	6
E	5	0	1	5
$\sum$	12 jel			30 bit

$$\langle d_k \rangle = \frac{30 bit}{12 jel} = 2,5 bit/jel$$

A bitsorozathoz tartozó üzenet: KEND␣LE

Ha egy forrásban olyen valószínűségekkel jönnének a jelek, mint a BENEDEK␣ELEK szóban az arányuk, mindig az előző jeltől függetlenül valószínűséggel (IRD forrás), akkor nem lehet az entrópiánál kevesebb átlagos bittel kódolni akárhogy trükközök.

Általános információelméleti eredmények: entrópia

k jel	n_k	p_k	$I_k = \log_2(1/p_k)$ [bit]	$p_k I_k$
␣	1	1/12	$\log_2 12/1 = 3,585$	0,2987
B	1	1/12	3,585	0,2987
D	1	1/12	3,585	0,2987
L	1	1/12	3,585	0,2987
N	1	1/12	3,585	0,2987
K	2	2/12	$\log_2 12/2 = 2,585$	0,4308
E	5	5/12	$\log_2 12/5 = 1,263$	0,5262
össz	$n = 12$			H=2,451 bit/jel

$$H = \sum_k p_k \log_2 \frac{1}{p_k} = \sum_k p_k I_k$$

$$H = 4 \cdot 0,2987 + 0,4308 + 0,5262 = 2,451 \text{ bit/jel}$$

Általános információelméleti eredmények: hatásfok

Az IRD forrás entrópiája

$$H = 2,451 \text{ bit/jel}$$

7 jel esetén a maximális entrópia (amikor minden jel 1/7 valószínűséggel jön)

$$H_{7,max} = \log_2 7 = 2,807 \text{ bit/jel}$$

A hatásfok:

$$\eta = \frac{H}{H_{7,max}} = 0,873 \approx 87\%$$

A redundancia

$$R = 1 - \eta \approx 13\%$$

Lackfi János: Parabola

Parabola, parabola
antenna,
nézzünk tévét
éppen ma!
Parabola, parabola
futbalmeccs,
lassított gól,
sípcsont reccs!
Parabola, parabola
sminkreklám.
Bőröd fittyed?
Kend ezt rá!

Parabola, parabola
bankrablók,
lő, fut, robban,
égből lóg.
Parabola, parabola
popcsillag,
rázós ritmus
popsidnak.
Parabola, parabola
antenna,
űrlény caplat
álmodba!

Adaptív Huffman-kódolás

A Huffman-kódolás esetén előre kellene tudnom, hogy milyen a forrás statisztikája. Gyakorlatban gyakran nem tehetem meg, hogy végigvárom a jelsorozatot, és csak utána kezdek el kódolni.

Sőt ez a statisztika időben változhat. Van olyan változata a Huffman-kódolásnak, amely a kódszavakat időben változtatva hozzáigazítja a közelmúltbeli statisztikához. Tehát a kódhosszak úgy változnak, hogy várhatóan egyre hatékonyabb lesz a kódolás.

Általában adaptív kódolásnak nevezzük az olyan kódolásokat, amelynél a forrás tulajdonságainak figyelembe vételével egyre hatékonyabban tudom kódolni a szöveget. A hatékonyabb alatt azt értem, hogy átlagosan kevesebb bittel jelenként.

Vázlat

Aritmetikai kód: kódolás

Az aritmetikai kódolásnál a [0; 1[intervallumot szűkítgetjük az egymás utáni jelekkel, míg adott pontosságú fixpontos ábrázolásban egyetlen szám marad benne.

A fixpontos szám alakja
$0,bbbbbbbb_2$
ahol a b-k bináris számjegyeket (1 vagy 0) jelölnek (itt például 8 darabot).
(A tizedesvessző előtti nullát felesleges tárolni, mivel mindig nulla.)

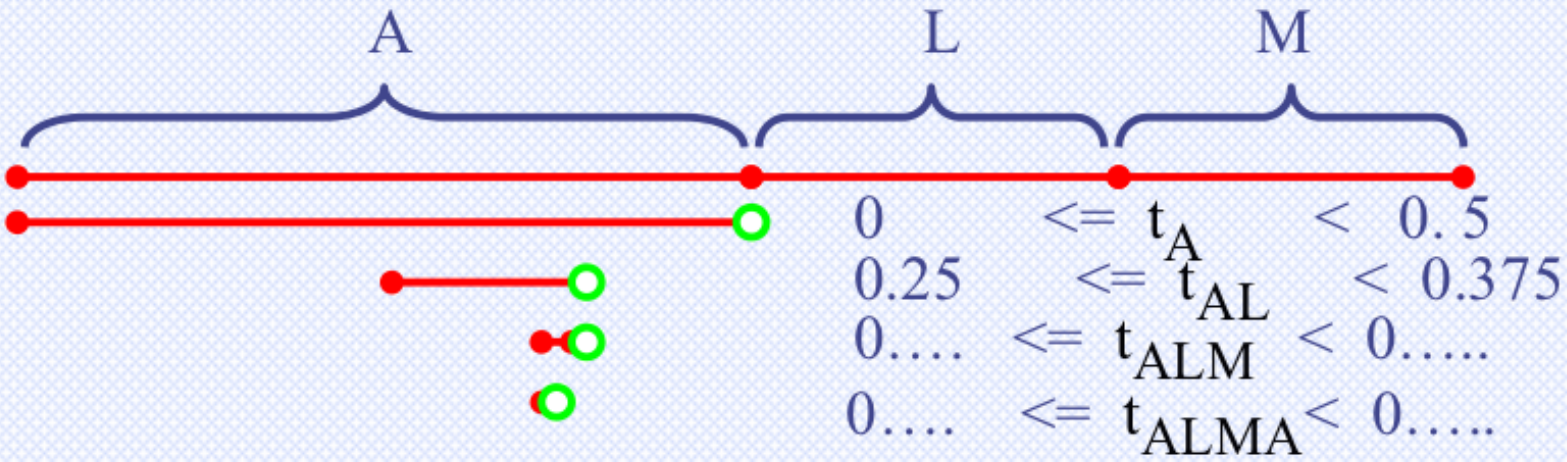
A nulla utáni 8 bináris számjegy esetén például a következő értékek tárolhatóak:

$$0, \quad \frac{1}{256}, \quad \frac{2}{256}, \quad \frac{3}{256}, \quad \ldots \quad \frac{254}{256}, \quad \frac{255}{256}$$

Egyszerű példa aritmetikai kódolásra

Kódolandó szöveg: **ALMA**

Szimbólumok	Előfordulások száma	Relatív gyakoriság	Számtartomány
A	2	2/4	$0 <= t_A < 0.5$
L	1	1/4	$0.5 <= t_L < 0.75$
M	1	1/4	$0.75 <= t_M < 1$



- A fenti példában az intervallumok hossza annak felel meg, hogy az adott üzenet milyen valószínűséggel áll elő, ha a forrásban a jelek az előző jelektől függetlenül mindig a fenti valószínűségekkel érkeznek (IRD forrás).
-

$$p_A = 1/2$$

$$p_{AL} = 1/2 \cdot 1/4 = 1/8$$

$$p_{ALM} = 1/2 \cdot 1/4 \cdot 1/4 = 1/32$$

$$p_{ALMA} = 1/2 \cdot 1/4 \cdot 1/4 \cdot 1/2 = 1/64$$

$$p_{LMLM} = 1/4 \cdot 1/4 \cdot 1/4 \cdot 1/4 = 1/256$$

- A továbbiakban a szűkítés során az üzenet feldolgozott szakaszának valószínűségét fogjuk p_üzenet-tel jelölni. Ha egy lépés során a k jel jön, akkor p_üzenet értéke a k jel valószínűségével fog szorzódni:
  `p_üzenet = p_üzenet * p[k]`

```
Be: üzenet
az üzenet k jeleinek ABC-sorrendbe rakása
p[k] valószínűségek meghatározása
a[k] alsó határok meghatározása
alsó = 0              # kezdőintervallum alsó határa
p_üzenet = 1          # kezdőintervallum hossza
Ciklus az összes k jelre az üzenetben:
    alsó = alsó + p_üzenet * a[k]
    p_üzenet = p_üzenet * p[k]
    felső = alsó + p_üzenet
    Ki: k, alsó, felső
Ciklus vége
kód = egy szám a legutolsó intervallumból.
```

Egyes jelek kódjai

A $[0;1[$ intervallumot osztjuk fel az egyes jelek között. A jeleket ABC-rendbe (adott kódlap szerinti rendbe) rakjuk, és mindegyiknek a valószínűségével egyező nagyságú intervallum jut az egységnyi hosszúságú intervallumon.

Az ALMA illetve MIKKAMAKKA üzenetek esetén itt láthatóak az egyes k jelekhez tartozó p_k valószínűségek és a_k alsó határok.

k	A	L	M
p_k	1/2	1/4	1/4
a_k	0	1/2	3/4

k	A	I	K	M
p_k	0,3	0,1	0,4	0,2
a_k	0	0,3	0,4	0,8

Jelsorozat kódja

Az aritmetikai kódolásnál a [0; 1[intervallumot szűkítgetjük az egymás utáni jelekkel, míg adott pontosságú fixpontos ábrázolásban egyetlen szám marad benne.

M	$\rightarrow$ [0,8	; 1	[
MI	$\rightarrow$ [0,86	; 0,88	[
MIK	$\rightarrow$ [0,868	; 0,876	[
MIKK	$\rightarrow$ [0,8712	; 0,8744	[
MIKKA	$\rightarrow$ [0,8712	; 0,87216	[
MIKKAM	$\rightarrow$ [0,871968	; 0,87216	[
MIKKAMA	$\rightarrow$ [0,871968	; 0,8720256	[
MIKKAMAK	$\rightarrow$ [0,87199104	; 0,87201408	[
MIKKAMAKK	$\rightarrow$ [0,872000256	; 0,872009472	[
MIKKAMAKKA	$\rightarrow$ [0,872000256	; 0,8720030208	[

MIKKAMAKKA$\rightarrow 0,872002$

Aritmetikai kód: kódolt üzenet visszafejtése

```
Be: jelek és valószínűségek (k, p[k])
a[k] alsó határok kiszámítása
Be: kód     # az üzenet kódja
Be: hossz   # az üzenet hossza
Ciklus hossz-szor:
    k = a jel, aminek az intervallumában van a kód
    Ki: k
    kód = (kód - a[k]) / p[k]
Ciklus vége
```

Visszafejtés

0,872002	M
0,360010	I
0,600100	K
0,500250	K
0,250625	A
0,835417	M
0,177083	A
0,590278	K
0,475694	K
0,189236	A

Egy gyakorlati alkalmazás

A képek JPEG kódolásánál választható az aritmetikai és Huffman-kódolás az adatok egyik tömörítési lépéséhez.
A Huffman-kódolás gyorsabb, de a tömörítés mértéke elmarad az aritmetikai kódoláshoz képest.