Informatika – Kódování

Radim Farana

Podklady předmětu Informatika
pro akademický rok 2007/2008

Obsah

- Základy pojmy z diskrétních kódů.
- Druhy kódů.
- Nejkratší kódy.
- Detekce chyb, Hammingova vzdálenost.
- Kontrolní a samoopravné kódy.
  - Lineární kódy.
  - Cyklické kódy.

Kód

- Popis přiřazení kódových slov jednotlivým zprávám (kódová kniha).
- Kódové slovo je posloupnost znaků použité abecedy.
- Abeceda je množina znaků (binární abeceda $Z_2 = \{0, 1\}$)
- Minimální délka kódového slova:
  $N^*(x) = - \log_2(P(x))$ [bit]

Vlastnosti kódu

- prosté kódování: různým zprávám odpovídají různá kódová slova,
- jednoznačná dekódovatelnost: ze znalosti zakódované zprávy lze jednoznačně určit zprávu zdrojovou,
- Kód K : A → B musí být prostým zobrazením.

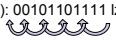
Problém dekódování

Zpráva				
Kód A	0	01	001	111

Posloupnost zpráv (kódových slov): 00100101111 nelze jednoznačně dekódovat

Kód B	0	01	011	111

Posloupnost zpráv (kódových slov): 00101101111 lze jednoznačně dekódovat?



Ano, ale jen „odzadu“, po přijetí celé posloupnosti zpráv.

Kód C	0	10	110	111

Posloupnost zpráv (kódových slov): 01011010111 můžeme dekódovat on-line.

Důvod? Žádné kódové slovo není začátkem jiného kódového slova (prefixem).

Druhy kódů

- Prefixový kód je prosté kódování u kterého žádné kódové slovo není začátkem jiného kódového slova.
- Blokový kód je prosté kódování u kterého mají všechna kódová slova stejnou délku (počet znaků). Protože musí být prostým zobrazením, je nutně také prefixovým kódem.

Použití kódů

- Nejkratší (optimální) kódy $R \rightarrow 0$, $L \rightarrow \min$,
- Bezpečnostní kódy
  - detekční kódy (odhalují chyby),
  - samoopravné kódy (opravují chyby),
- Speciální kódy
  - kódy konstantní změny (Grayův kód),
  - čárové kódy,
  - alfanumerické kódy,
  - číselné kódy (datové formáty), …

Nejkratší kódy

- Pokud má $R \rightarrow 0$, neboli $L \rightarrow \min$, pak
- $N(i) \rightarrow N^*(i)$ pro $i = 1, 2, \ldots n$.
- Hledáme vhodný algoritmus konstrukce takového kódu:
  - Huffmanův kód (1952),
  - Shannonův kód.

Algoritmy se liší, stejně tak i dosažené výsledky, Huffmanův kód se snáze algoritmizuje a tedy také realizuje

Huffman, David A.

* 1925, USA
+ 7. 10. 1999 California, USA
http://www.ucsc.edu/currents/99-00/10-11/huffman.html

Huffmanův kód

Triviální případ

Zpráva	$P(i)$	kód
1	>	0
2	<	1

Redukovaná abeceda

Zpráva	$P(i)$	redukce	kód	expanze
1	»	1	0	0
2	>	2,3	1	10
3	<			11

Příklad

Zpráva	A	B	C	D	E
$P(i)$	0,4	0,3	0,1	0,1	0,1
1.redukce	0,4	0,3	0,2	0,1	
2.redukce	0,4	0,3	0,3		
3.redukce	0,6	0,4			
znaky	0	1			
kód	1	00	011	0100	0101

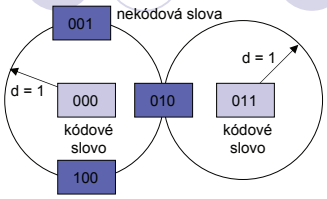
Postup:
- seřazení podle pravděpodobnosti,
- postupná redukce a oprava pořadí,
- přiřazení znaků 0, 1 a zpětná expanze.

Problémy:
- definice pořadí zpráv pro stejnou $P(i)$,
- zařazení skupin se stejnou $P(i)$,
- pořadí přiřazení znaků 0, 1.

Detekce chyb

- Množinu všech slov rozdělíme na slova kódová a slova nekódová.
- *t*-násobná chyba změní kódové slovo na nekódové, pokud se dvě kódová slova liší ve více než *t* znacích.
- Hammingova vzdálenost je počet znaků ve kterých se dvě kódová slova liší.
- Hammingova vzdálenost kódu *d* je nejmenší z nich.

Hammingova vzdálenost



Hamming, Richard Wesley

* 11. 2. 1915 Chicago, Il. USA
† 7. 1. 1998 Monterey, Cal. USA
http://cm.bell-labs.com/cm/cs/alumni/hamming/

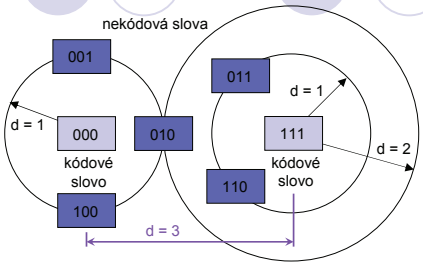
Kód odhaluje *t*-násobné chyby, pokud je Hammingova vzdálenost kódu $d > t$.

Označení kódů (blokových): (n, k)-kód

počet znaků / počet informačních znaků

(4, 3)-kód má jeden kontrolní znak, je schopen mít $d = 2$.

Opravování chyb



Kód opravuje *t*-násobné chyby, pokud je Hammingova vzdálenost kódu $d > 2.t$.

Lineární kódy (maticové kódy)

- Kódové slovo chápeme jako řádkový vektor **v** = [0 0 1].
- Lineární kombinací libovolného počtu kódových slov vznikne opět kódové slovo.
- Kód je možno popsat pomocí generující matice (kterou tvoří báze kódu).

$$\mathbf{G} = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \qquad \mathbf{v} = \mathbf{z}.\mathbf{G}$$

Systematické kódy

- Informační slovo tvoří začátek kódového slova. $\mathbf{G} = [\,\mathbf{E} \mid \mathbf{B}\,]$
- Určení informačního slova (dekódování) je triviální.
- Je možno snadno určit kontrolní matici kódu $\mathbf{H} = [\,-\mathbf{B}^T \mid \mathbf{E}'\,]$
- A syndrom přijatého slova $\mathbf{s} = \mathbf{H}.\mathbf{v}^T$
- Nenulový syndrom indikuje chybu.

Hammingovy kódy

- Opravují jednoduché chyby a
- jsou perfektní = při daných vlastnostech mají minimální možnou redundanci.
- Kód s m kontrolními znaky (m = 2, 3, …) má délku $n = 2^m - 1$.
- Sloupce kontrolní matice tvoří binární rozvoj čísel 1, 2, …, $2^m - 1$
- Nenulový syndrom je binárním rozvojem pozice chyby.

Cyklické kódy (polynomické kódy)

- Jsou podtřídou lineárních kódů. Kódové slovo chápeme jako zápis polynomu.
- Cyklickým posunem znaků kódového slova vznikne opět kódové slovo.
- Kromě generující matice mohou být popsány také generujícím polynomem.
- Jsou schopny dobře detekovat (opravovat) i shlukové chyby.

Realizace cyklických kódů

- Informační slovo dělíme generujícím polynomem,
- určíme zbytek po dělení,
- zbytek připojíme za informační slovo.
- Celé kódové slovo je dělitelné generujícím polynomem beze zbytku.
- Pod označením CRC-kódy (Cyclic Redundance Code) mají široké použití

Typické CRC kódy

počet kontrolních bitů	označení	generující polynom	použití
8	LRCC–8	z^8+1	kontrolní Byte je součet datových Byte modulo 2
12	CRC–12	$z^{12}+z^{11}+z^3+z^2+z+1$	používá se pro šestibitové znaky
16	LCRC–16	$z^{16}+1$	kontrolní součet dvojic Byte (Word) modulo 2
16	CRC–16	$z^{16}+z^{15}+z^2+1$	binární synchronní protokol
16	CRC–16 reverzní	$z^{16}+z^{14}+z+1$	
16	SDLC	$z^{16}+z^{12}+z^5+1$	linkový protokol IBM, standard CCITT
16	SDLC reverzní	$z^{16}+z^{11}+z^4+1$	
32	CRC–32	$z^{32}+z^{26}+z^{23}+z^{22}+$ $+z^{16}+z^{12}+z^{11}+z^{10}+$ $+z^8+z^7+z^5+z^4+$ $+z^2+z+1$	Ethernet, HDLC, ZMODEM