

Josef Pužman

Poznámka ke kódování psané češtiny

Kybernetika, Vol. 2 (1966), No. 1, 86--87

Persistent URL: <http://dml.cz/dmlcz/125328>

Terms of use:

© Institute of Information Theory and Automation AS CR, 1966

Institute of Mathematics of the Academy of Sciences of the Czech Republic provides access to digitized documents strictly for personal use. Each copy of any part of this document must contain these

Terms of use.



This paper has been digitized, optimized for electronic delivery and stamped with digital signature within the project *DML-CZ: The Czech Digital Mathematics Library*
<http://project.dml.cz>

Poznámka ke kódování psané češtiny

Ve svém článku [1] se J. Kraus dotýká mj. problému optimálního kódování grafémů české abecedy, tj. konstrukce nerovnoměrného dvojkového kódu s vlastností prefixu pro zadaný systém $\{x_i\}$ relativních četností výskytu grafémů (např. z [3]) tak, aby průměrná délka N kódových posloupností byla co nejmenší. Používá k tomu Shannonovy-Fanovy metody [2], která však vždy nezaručuje optimální kód; zaručuje jen splnění vlastností prefixu. Poslední znamená, že jednotlivé kódové posloupnosti jsou od sebe rozlišitelné v tom smyslu, že žádná kódová posloupnost není rozšířením jiné, čili žádná kódová posloupnost není obsažena v jiné jako její prefix. Význam rozlišitelnosti pak spočívá v tom, že kódové posloupnosti lze řadit ve zprávy (věty) bez rozdělovacích znaků, přičemž dekódování takové zprávy je jednoznačné.

Kód uvedený v [1] však vlastnost prefixu nespĺňuje. Je možné nalézt, že grafém z (0100) je prefixem dvou grafémů: d (01001) a j (01000), takže např. větu „010011101000011“ lze dekódovat buď jako *zed* (0100 11101 000011) nebo jako *dař* (01001 11010 00011).

Vyžadujeme-li od kódu, aby byl ještě optimální, nelze předem klást omezení na délky kódových posloupností, neboť ty závisí též na systému relativních četností $\{x_i\}$. Minimální průměrnou délku N lze docílit přiřazením co nejkratších kódových posloupností symbolům s největší četností na úkor symbolů s malou četností. Bez ohledu na $\{x_i\}$ leží maximální délka kódových posloupností optimálního dvojkového kódu mezi $\log_2 k$ a $k - 1$, kde k je počet kódovaných symbolů. Při kódování grafémů české abecedy, jichž je 42 (včetně mezery mezi slovy), může být maximální délka až 41 (pro skutečný systém četností je však menší).

U optimálních kódů je snahou docílit průměrnou délku kódových posloupností, kterou označíme N , co nejmenší a protože (pro dvojkový nerovnoměrný kód) N nemůže být menší než $-\sum_{i=1}^k x_i \log_2 x_i$, což je, za předpokla-

Tab. 1. Optimální nerovnoměrný kód grafémů české abecedy

Pís- meno	Četnost	Kódová posloupnost	Vážená délka
mezera	0,16586	000	0,49758
e	0,07261	0010	0,29044
o	0,06766	0011	0,27064
a	0,05431	0100	0,21724
n	0,04036	0101	0,16144
v	0,03953	01100	0,19765
t	0,03870	01101	0,19350
s	0,03743	01110	0,18715
k	0,03368	01111	0,16840
i	0,03303	10000	0,16515
l	0,03293	10001	0,16465
u	0,02999	10010	0,14995
r	0,02932	10011	0,14660
p	0,02792	10100	0,13960
m	0,02788	10101	0,13940
d	0,02643	10110	0,13215
í	0,02490	10111	0,12540
á	0,02153	11000	0,10765
j	0,02088	11001	0,10440
z	0,01902	110100	0,11412
y	0,01623	110101	0,09738
ň	0,01437	110110	0,08622
b	0,01363	110111	0,08178
h	0,01095	111000	0,06570
é	0,01046	111001	0,06276
c	0,01045	111010	0,06270
ch	0,00974	1110110	0,06818
ř	0,00971	1110111	0,06797
ž	0,00956	1111000	0,06692
ý	0,00852	1111001	0,05964
č	0,00784	1111010	0,05488
š	0,00746	1111011	0,05222
ř	0,00652	1111100	0,04564
ě	0,00619	1111101	0,04333
ú, ů	0,00546	1111110	0,03822
ď	0,00414	11111110	0,03312
f	0,00194	111111110	0,01746
g	0,00168	1111111110	0,01680
x	0,00062	11111111110	0,00682
ó	0,00042	111111111110	0,00504
w	0,00011	1111111111110	0,00143
q	0,00003	1111111111111	0,00039
	1,00000		4,70681

du nezávislosti, entropie na symbol H kódované abecedy, je snahou docílit, aby se N lišilo co nejméně od H . Mírou účinnosti nerovnoměrného kódu je poměr $H/N \leq 1$ (udávaný někdy v procentech). Podle [3] je $H = 4,6665$ bitů/symbol a protože v [1] bylo nalezeno $N = 5,013$, je účinnost kódu 93,1%. Rozdíl mezi hodnotou $H = 4,6665$ a hodnotou z [1] $H = 4,96$ je způsoben zaokrouhlením některých četností a neuvažováním grafémů \acute{e} , q a w , takže v úvahách se používá neúplný

systém relativních četností ($\sum_{i=1}^{39} x_i = 0,988 < 1$). Použitím Fanovy metody kódování dosáhneme průměrnou délku N kódových posloupností 4,71635, takže účinnost vzroste na 98,94%, ale ani to ještě není optimální kód.

Skutečně optimální kód lze sestavit Huffmanovu metodu [4]. V tab. 1 je uvedeno 42 grafémů české abecedy seřazených podle klesajících relativních četností, které byly zjištěny Ústavem pro jazyk český ČSAV (slabiky $d\acute{e}$, te , $n\acute{e}$ byly považovány za $d\acute{e}$, te , $n\acute{e}$) [3] a optimální nerovnoměrný kód s vlastností pre-

fixu. Průměrná délka N kódových posloupností je 4,70681 a účinnost je skutečně dobrá: 99,14%. Účinnost lze pak zlepšit jen kódováním dvojic, trojic, atd. grafémů do nerovnoměrného kódu s ohledem na závislosti skutečného textu.

(1. VII. 1965)

Josef Pužman

LITERATURA

- [1] Kraus J.: Kódování a komprese psané češtiny. *Kybernetika 1* (1965), č. 1, 74–84.
- [2] Fañó R. M.: *Transmission of information*. J. Wiley, New York 1961.
- [3] Doležel L.: Předběžný odhad entropie a redundance psané češtiny, *Slovo a slovesnost XXIV* (1963), č. 3, 165–175.
- [4] Huffman D. A.: A method for the construction of minimum-redundancy code. *Proc. IRE 40* (1952), č. 9, 1098–1101.