

Plan wykładu	Teoria informacji	Kodowanie	Znaki alfanumeryczne	Pytania
	○○○ ○○○	○○○○○ ○○○○○	○ ○ ○○○○ ○○○	

Kodowanie informacji

Kodowanie znaków

prof. dr hab. inż. Joanna Józefowska

Poznań, rok akademicki 2008/2009

prof. dr hab. inż. Joanna Józefowska Kodowanie informacji

Plan wykładu	Teoria informacji	Kodowanie	Znaki alfanumeryczne	Pytania
	○○○ ○○○	○○○○○ ○○○○○	○ ○ ○○○○ ○○○	

Plan wykładu

- 1 Teoria informacji
 - Pojęcie informacji
 - Ilość informacji
- 2 Kodowanie
 - Założenia
 - Jednostki informacji
- 3 Znaki alfanumeryczne
 - Kod EBCDIC
 - Kod ASCII
 - UNICODE
- 4 Pytania

prof. dr hab. inż. Joanna Józefowska Kodowanie informacji

Plan wykładu	Teoria informacji	Kodowanie	Znaki alfanumeryczne	Pytania
	●○○ ○○○	○○○○○ ○○○○○	○ ○ ○○○○ ○○○	

Informacja → wiadomość → komunikat

Komunikat

jest odpowiednio zakodowaną wiadomością zawierającą pewną ilość informacji.

Przykład

- **Wiadomość:** Rozpoczęto zbrojne wystąpienie przeciwko rządowi republikańskiemu w Hiszpanii.
- **Komunikat:** "Nad Hiszpanią niebo jest czyste."
- **Informacja:** Rozpoczęła się wojna domowa w Hiszpanii.

prof. dr hab. inż. Joanna Józefowska Kodowanie informacji

Plan wykładu	Teoria informacji	Kodowanie	Znaki alfanumeryczne	Pytania
	●○○ ○○○	○○○○○ ○○○○○	○ ○ ○○○○ ○○○	

Informacja

Informacja (A. Mazurkiewicz)

wielkość abstrakcyjna, która może być przechowywana w pewnych obiektach, przesyłana pomiędzy obiektami, przetwarzana w pewnych obiektach i stosowana do sterowania pewnymi obiektami, przy czym przez obiekty rozumie się organizmy żywe, urządzenia techniczne oraz systemy takich obiektów.

prof. dr hab. inż. Joanna Józefowska Kodowanie informacji

Teoria informacji

Teoria informacji

dział matematyki na pograniczu statystyki i informatyki, mający również olbrzymie znaczenie w współczesnej telekomunikacji, dotyczący przetwarzania informacji oraz jej transmisji, kompresji, kryptografii itd.

Entropia informacyjna

Średnia ważona ilość informacji (H) w komunikatach ze źródła wytwarzającego n różnych komunikatów z prawdopodobieństwem $p_i, i = 1, \dots, n$ każdy, określana jest wzorem:

$$H = \sum_{i=1}^n p_i \log_2 \frac{1}{p_i}$$

Średnia ilość informacji jest nazywana **entropią informacyjną**. Intuicyjnie entropia źródła jest to wartość oczekiwana ilości informacji w komunikatach tego źródła.

Zawartość informacyjna wiadomości

Dla odniesienia zawartości informacyjnej (k) do prawdopodobieństwa, Shannon wprowadził proste równanie:

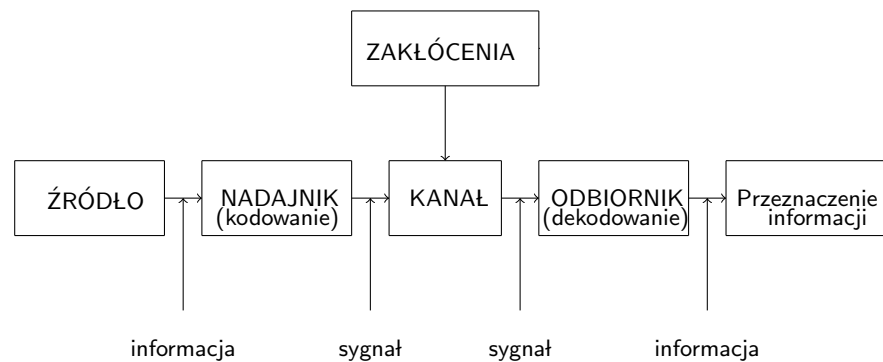
$$k = \log_2 \frac{1}{p}$$

gdzie p jest prawdopodobieństwem wystąpienia wiadomości.

Potrzeba kodowania informacji

lista	dokument	arkusz kalkulacyjny	zdjęcie	rysunek	dźwięk	obraz wideo
plik danych	plik tekstowy	plik arkusza kalkulacyjnego	plik z grafiką rastrową	plik z grafiką wektorową	plik dźwiękowy	plik wideo
rekordy pola	strony słowa	wiersze i kolumny	bitmapy	wektory (listy do wyświetlania)	głos w postaci cyfrowej	wideo bitmapy
bajty						
bity						

Przesyłanie informacji



Zasady kodowania

Kodowanie

jest to przyporządkowanie znakom jakiegoś alfabetu ciągów znaków innego alfabetu.

- na przykład: $a \leftarrow 00110$, $b \leftarrow 01010$,
- inaczej: $K(a) = 00110$, $K(b) = 01010$,
- $K(a)$ jest **słowem kodowym** a , czyli 00110 jest słowem kodowym a ,
- po zakodowaniu słowa x do postaci y można je odkodować tylko na jeden sposób uzyskując x ,
- kodowanie binarne każdemu znakowi alfabetu przypisuje ciąg binarny,
- jeżeli wszystkie słowa kodowe mają tę samą długość, to mówimy o kodowaniu **o stałej długości**, (np. ASCII).

Cele kodowania

- zmniejszenie objętości danych – kompresja,
- zapewnienie odporności na błędy – kody korekcyjne,
- zapewnienie poufności danych – kryptografia.

Redundancja

Średnia ważona długość słowa kodowego (L) dla źródła wytwarzającego n różnych komunikatów określana jest wzorem:

$$L = \sum_{i=1}^n p_i d_i$$

gdzie: p_i – prawdopodobieństwo wystąpienia i -tego komunikatu, d_i – długość słowa kodowego i -tego komunikatu (w bitach), $i = 1, \dots, n$.

Redundancja

Redundancją danego sposobu kodowania dla danego źródła nazywa się różnicę wartości: $L - H$.

Twierdzenie Shannona

Twierdzenie

Redundancja jest liczbą nieujemną.

Redundancja

- Redundancja to ilość informacji przekraczająca wymagane do rozwiązania problemu minimum.
- Nadmiarowość jest cechą każdego systemu informacyjnego przesyłającego jakieś dane cyfrowe.
- Ułatwia odtworzenie danych po ich częściowej utracie, czy uszkodzeniu lub też wykrycie takiego uszkodzenia (bit parzystości, suma kontrolna, kod nadmiarowy CRC)
- Usuwanie nieprzydatnej redundancji to kompresja danych.

Bit

- **bit** – podstawowa, najmniejsza i niepodzielna jednostka informacji cyfrowej, jaka może być przetwarzana przez komputer,
- bit przyjmuje wartości oznaczane jako 0 albo 1,
- bit może przechowywać informacje o wystąpieniu jednego z dwóch możliwych stanów,
- bit to skrót terminu **B**inary **D**igi**T**,
- bit to po angielsku *kawałek*,
- skrót: **b**.

Liczba bitów a kodowanie

liczba bitów	można zakodować	przykład
1 bit	2 znaki	0 i 1
2 bity	4 znaki	00, 01, 10 i 11
3 bity	8 znaków	000, 001, 010, 011, 100, 101, 110 i 111
4 bity	16 znaków	0000, ..., 1111
8 bitów	256 znaków	

Bloki bitów

liczba bitów	nazwa bloku
4 bity	tetrada
8 bitów	bajt (oktet)
2 bajty	półsłowo
4 bajty	słowo
64 bity	podwójne słowo

Słowo

- Słowo komputerowe - ilość informacji, na której operuje komputer. Komputer 8-, 16-, 32- (64-, 128-) bitowy oznacza wielkość grupy danych, którą komputer może operować jako całością.
- Słowo procesora - jednostka informacji o długości naturalnej dla danego procesora (długość odpowiada długości rejestrów - obecnie 32 lub 64 bity).
- Słowo pamięci - jednostka informacji możliwa do przetransmitowania w jednym cyklu transmisji do lub z pamięci (obecnie zwykle 64 bity, niekiedy 128).

Rozmiar danych

liczba bajtów	nazwa	przykład
$2^{10} = 1024$	kilobajt [KB]	typowa strona tekstu to kilka KB
$2^{20} = 1024K$	megabajt [MB]	minuta muzyki
$2^{30} = 1024M$	gigabajt [GB]	film cyfrowy, ludzki genom
$2^{40} = 1024G$	terabajt [TB]	duża biblioteka
$2^{50} = 1024T$	petabajt [PB]	ludzka pamięć

Dlaczego komputery przetwarzają bity?

- za pomocą bitów można wygodnie kodować dowolną informację,
- bity są proste – mogą przyjmować tylko jeden z dwóch stanów: 0 i 1,
- stosunkowo łatwo zbudować urządzenia rozpoznające dwa stany (napiecie elektryczne),
- bity są odporne na zakłócenia - w czasie przekazu należy wykryć tylko dwa poziomy, wysoki H - 1 i niski L - 0,
- binarny system pozycyjny da się wprost zakodować za pomocą bitów - każdej cyfrze binarnej odpowiada jeden bit (na bitach można wykonywać dowolne operacje arytmetyczne, a więc liczyć).

Kod EBCDIC

- EBCDIC - Extended Binary Coded Decimal Interchange Code,
- 8-bitowe kodowanie znaków stworzone jako rozszerzenie kodowania BCD,
- umożliwia zapisanie do 256 różnych symboli,
- pojedynczy znak zapisywany jest na 8-bitach podzielonych na dwie części,
- pierwsze 4 bity oznaczają grupę do jakiej dany znak należy, a ostatnie 4 bity identyfikują konkretny znak.

Kod EBCDIC c.d.

- Kodowanie EBCDIC było używane głównie w systemach IBM (komputery: z/OS, VM, OS/390, systemy operacyjne: OS/400, i5/OS) w latach 60-tych XX wieku,
- ułatwiało wprowadzanie danych do komputera przy użyciu kart perforowanych,
- kody EBCDIC nie są zgodne z ASCII,
- EBCDIC występowało w wielu wersjach, odmiennych dla różnych państw,
- większość stron kodowych EBCDIC dopuszczała stosowanie tylko dwóch języków (angielski + dodatkowy).

Potrzeba kodowania

- do przechowywania i przetwarzania danych przez układy elektroniczne komputera używany jest system binarny,
- należy przedstawić tekst za pomocą liczb czyli jednoznacznie przyporządkować literom i innym znakom alfanumerycznym - liczby (numery),
- w ten sposób powstał w 1965 r. kod **ASCII (American Standard Code for Information Interchange)**.

Kod ASCII

- kod jest jawny i używany przez wszystkich użytkowników i twórców oprogramowania,
- jest to kod 7 bitowy, a więc możemy za jego pomocą przedstawić 2^7 czyli 128 znaków,
- liczby 0-31 oraz 127 odpowiadają znakom sterującym takim jak wysunięcie strony, tabulacja itp.,
- w 1981 r. IBM wprowadził rozszerzony do 8 bitów kod, co pozwala na przedstawienie 256 znaków (w tym znaki specjalne, graficzne, matematyczne i diakrytyczne znaki narodowe),
- interpretacja znaków o kodach od 128 do 255 zależy od zastosowanej tzw. strony kodowej (np. ISO Latin-2 dla języka polskiego).

Kod ASCII

Kod ASCII

Dec	Hex	Znak	Skrót	Dec	Hex	Znak	Skrót
0	00	Null	NUL	16	10	Data Link Escape	DLE
1	01	Start Of Heading	SOH	17	11	Device Control 1 (XON)	DC1
2	02	Start of Text	STX	18	12	Device Control 2	DC2
3	03	End of Text	ETX	19	13	Device Control 3 (XOFF)	DC3
4	04	End of Transmission	EOT	20	14	Device Control 4	DC4
5	05	Enquiry	ENQ	21	15	Negative Acknowledge	NAK
6	06	Acknowledge	ACK	22	16	Synchronous Idle	SYN
7	07	Bell	BEL	23	17	End of Transmission Block	ETB
8	08	Backspace	BS	24	18	Cancel	CAN
9	09	Horizontal Tab	HT	25	19	End of Medium	EM
10	0A	Line Feed	LF	26	1A	Substitute	SUB
11	0B	Vertical Tab	VT	27	1B	Escape	ESC
12	0C	Form Feed	FF	28	1C	File Separator	FS
13	0D	Carriage Return	CR	29	1D	Group Separator	GS
14	0E	Shift Out	SO	30	1E	Record Separator	RS
15	0F	Shift In	SI	31	1F	Unit Separator	US

Dec	Hex	Znak	Dec	Hex	Znak	Dec	Hex	Znak	Dec	Hex	Znak
32	20	SPC	33	21	!	34	22	"	35	23	#
36	24	\$	37	25	%	38	26	&	39	27	'
40	28	(	41	29	)	42	2A	*	43	2B	+
44	2C	,	45	2D	-	46	2E	.	47	2F	/
48	30	0	49	31	1	50	32	2	51	33	3
52	34	4	53	35	5	54	36	6	55	37	7
56	38	8	57	39	9	58	3A	:	59	3B	;
60	3C	i	61	3D	=	62	3E	¿	63	3F	?
64	40	@	65	41	A	66	42	B	67	43	C
68	44	D	69	45	E	70	46	F	71	47	G
72	48	H	73	49	I	74	4A	J	75	4B	K
76	4C	L	77	4D	M	78	4E	N	79	4F	O
80	50	P	81	51	Q	82	52	R	83	53	S
84	54	T	85	55	U	86	56	V	87	57	W
88	58	X	89	59	Y	90	5A	Z	91	5B	[
92	5C	\	93	5D	]	94	5E	^	95	5F	_
96	60	`	97	61	a	98	62	b	99	63	c
100	64	d	101	65	e	102	66	f	103	67	g
104	68	h	105	69	i	106	6A	j	107	6B	k
108	6C	l	109	6D	m	110	6E	n	111	6F	o
112	70	p	113	71	q	114	72	r	115	73	s
116	74	t	117	75	u	118	76	v	119	77	w
120	78	x	121	79	y	122	7A	z	123	7B	{
124	7C	—	125	7D	g	126	7E	}	127	7F	Del

Kodowanie polskich znaków - heksadecymalnie

Kodowanie polskich znaków

Standard	ą	ć	ę	ł	ń	ó	ś	ź	ż	Ą	Ć	Ę	Ł	Ń	Ó	Ś	Ż	Ž
ISO 8859-21	B1	E6	EA	B3	F1	F3	B6	BC	BF	A1	C6	CA	A3	D1	D3	A6	AC	AF
CP 12502	B9	E6	EA	B3	F1	F3	9C	9F	BF	A5	C6	CA	A3	D1	D3	8C	8F	AF
CP 8523	A5	86	A9	88	E4	A2	98	AB	BE	A4	8F	A8	9D	E3	E0	97	8D	BD
Mazovia	86	8D	91	92	A4	A2	9E	A6	A7	8F	95	90	9C	A5	A3	98	A0	A1
Cyfromat	90	91	92	93	94	95	96	98	97	80	81	82	83	84	85	86	88	87
IEA-Swierk	A0	9B	82	9F	A4	A2	87	A8	91	8F	80	90	9C	A5	99	EB	9D	92
Ventura	96	94	A4	A7	91	A2	84	82	87	99	A5	A6	92	8F	8E	90	80	80
ELWRO-Junior	E1	E3	E5	EC	EE	EF	F3	FA	F9	C1	C3	C5	CC	CE	CF	D3	DA	D9
AmigaPL	E2	EA	EB	EE	EF	F3	F4	FA	FB	C2	CA	CB	CE	CF	D3	D4	DA	DB
TeXPL	A1	A2	A6	AA	AB	F3	B1	B9	BB	81	82	86	8A	8B	D3	91	99	9B
Atari-Calamus	D1	D2	D3	D4	D5	D6	D7	D8	D9	C1	C2	C3	C4	C5	C6	C7	C8	C9
CorelDraw!	E5	EC	E6	C6	F1	F3	A5	AA	BA	C5	F2	C9	A3	D1	D3	FF	E1	E
Unicode	105	107	119	142	144	F3	15B	17A	17C	104	106	118	141	143	D3	15A	179	17B

- Tekst zapisany w standardzie ISO – 8859-2:
Ą Ć Ę Ł Ń Ó Ś Ž Ż
ą ć ę ł ń ó ś ż ž
- Tekst wyświetlony w edytorze systemu Windows (Windows – 1250):
~ Ć Ę Ł Ń Ó — ˆ Ž
± ć ę ł ń ó ś Ł ž

Windows 1250

- Strona kodowa używana przez system Microsoft Windows do reprezentacji tekstów w językach środkowoeuropejskich używających alfabetu łacińskiego, takich jak albański, chorwacki, czeski, polski, rumuński, słowacki, węgierski,
- jest podobny do ISO 8859-2 — posiada wszystkie jego drukowalne znaki (a także kilka dodatkowych), lecz kilka z nich zajmuje inne miejsca.

UNICODE

- 256 znaków alfanumerycznych nie dawało możliwości zakodowania znaków diakrytycznych wielu języków np.: japońskiego, arabskiego, hebrajskiego itp.,
- odpowiedzią jest kod nazywany UNICODE o długości 16 bitów dla każdego znaku, a to daje już możliwość zakodowania 2^{16} czyli 65536 znaków,
- UNICODE jest jednoznaczny i uniwersalny, jest standardem dla XML,
- UNICODE mówi jakim znakom odpowiadają kody, a nie jak te znaki mają wyglądać.

Standardy UNICODE

- Standard Unicode definiuje kody numeryczne przypisane poszczególnym znakom, nie określa natomiast sposobu bajtowego kodowania znaków.
- Kodowanie określa sposób w jaki znaki ze zbioru mają być zapisane w postaci binarnej.
- Stosowane metody kodowania:
 - UCS - Universal Character Set,
 - UTF - Unicode Transformation Format.
- Niektóre z metod kodowania:
 - UTF-8 kodowanie 8-bitowe ze zmienną długością kodowania, zgodne z ASCII,
 - UTF-16 kodowanie 16-bitowe ze zmienną długością kodowania,
 - UTF-32/UCS-4 kodowanie 32-bitowe,
 - UTF-EBCDIC kodowanie 8-bitowe ze zmienną długością kodowania, zgodne z EBCDIC - nie jest częścią standardu.

Standardy UNICODE

- UTF - 8 - unikod ośmiobitowy; pierwsze 128 znaków pokrywa się z tablicą ASCII i jest zapisywana za pomocą jednego bajta, natomiast znaki dodatkowe (np. polskie literki) są zapisywane za pomocą specjalnych kilkubajtowych sekwencji.
- UTF -16 - kodowanie dwubajtowe; każdemu znakowi odpowiadają dwa bajty.
- UTF -32 - pewnym problemem mogą być języki ideograficzne (CJK), na przykład chiński liczący tysiące znaków. Problemy takie ostatecznie rozwiązuje kodowanie czterobajtowe UTF-32 (cztery bajty dają $2^{32} = 4$ miliardy kombinacji).

Plan wykładu	Teoria informacji ○○○ ○○ ○	Kodowanie ○○○○○ ○○○○○	Znaki alfanumeryczne ○ ○ ○○○○ ○○○	Pytania
--------------	----------------------------------	-----------------------------	--	---------

Wprowadzanie znaków w UNICODE

Typowe układy klawiatur udostępniają tylko niewielki zbiór znaków, na przykład standardowe polskie układy klawiatury zawierają wyłącznie polskie litery i znak euro, nie zawierają natomiast znaków cudzoalfabetycznych. Do wprowadzania znaków nieobecnych na klawiaturze można użyć jednej z poniższych metod:

- W systemie Windows, Mac OS X (od wersji 10.2) oraz w licznych dystrybucjach Linuksa (ze środowiskiem graficznym) dostępna jest tablica znaków, w której można wyszukać potrzebny znak, skopiować go do schowka, a następnie wkleić w program. W Systemach Windows 95, 98 i ME tablica znaków nie daje dostępu do wszystkich znaków Unikodu.

prof. dr hab. inż. Joanna Józefowska

Kodowanie informacji

Plan wykładu	Teoria informacji ○○○ ○○ ○	Kodowanie ○○○○○ ○○○○○	Znaki alfanumeryczne ○ ○ ○○○○ ○○○	Pytania
--------------	----------------------------------	-----------------------------	--	---------

Wprowadzanie znaków w UNICODE

Znając numer pozycji znaku (czy to dziesiętny, czy to szesnastkowy) można wprowadzić ten znak na kilka sposobów, nie wszystkie programy obsługują wszystkie poniższe sposoby:

- Trzymając naciśnięty (lewy) Alt i wpisując z klawiatury numerycznej numer dziesiętny tego znaku, po czym zwolnić klawisz Alt.
- Wpisać numer szesnastkowy tego znaku, a następnie wcisnąć kombinację (lewy) Alt+x. Przed numerem szesnastkowym nie może być cyfra ani litera z zakresu a ÷ f (gdyż zostaną uznane jako część kodu, ewentualnie należy poprzedzić kod np. spacją lub uzupełnić kod wiodącymi zerami do sześciu pozycji).
- Trzymając naciśnięty (lewy) Alt nacisnąć klawisz plus, wpisać szesnastkowy kod znaku, po czym zwolnić klawisz Alt. (Działa w systemie Windows XP i nowszych.)
- Trzymając naciśnięte klawisze Ctrl+Shift wpisać szesnastkowy numer znaku, a następnie puścić klawisze. (Działa w środowisku GNOME.)

prof. dr hab. inż. Joanna Józefowska

Kodowanie informacji

Plan wykładu	Teoria informacji ○○○ ○○ ○	Kodowanie ○○○○○ ○○○○○	Znaki alfanumeryczne ○ ○ ○○○○ ○○○	Pytania
--------------	----------------------------------	-----------------------------	--	---------

Pytania

- 1 Co to jest bit?
- 2 Co to jest kod ASCII?
- 3 Co to jest bajt?
- 4 Ile znaków można zakodować na ośmiu bitach?
- 5 Co to jest UNICODE?

prof. dr hab. inż. Joanna Józefowska

Kodowanie informacji