UNIVERSIDAD DE MÁLAGA

E.T.S. INGENIERÍA INFORMÁTICA

LENGUAJES Y CIENCIAS DE LA COMPUTACIÓN
UNIVERSIDAD DE MÁLAGA

Tema 1. Codificación y sus usos

José A. Montenegro

Dpto. Lenguajes y Ciencias de la Computación
ETSI Informática. Universidad de Málaga
monte@lcc.uma.es twitter

26 de septiembre de 2013

1. Mensajes
   - Definición Informal
2. Codificación
   - Aproximación al problema
   - Razones codificación mensajes
3. Definiciones Formales
   - Alfabetos, Símbolos, Mensajes y Cadena
   - Código, Palabra Codificada
   - Concatenación
   - Unívocamente Decodificable
4. Otros Alfabetos y Codificación
   - Código Morse
   - Código ASCII
   - Codificación PEM
   - Codificación de datos 2D
   - Ejemplo Doble Codificación

Mensajes

La primera tarea que debemos realizar es establecer el modelo matemático de un mensaje.

Ejemplo 1

Muchos mensajes están escritos en un lenguaje natural, como es el caso del Español.
Estos mensajes contienen símbolos y los símbolos forman palabras, los cuales constituyen frases.

Los mensajes pueden ser enviados de una persona a otra utilizando varios medios: una nota manuscrita, un correo electrónico, SMS, twitts, etc.

Ejemplo 2

Dispositivos como los escáners y cámaras digitales producen mensajes utilizando impulsos electrónicos. Estos mensajes pueden ser enviados de un dispositivo a otro mediante cables, o mediante ondas de radio.

Las definiciones formales basadas en estos ejemplos serán establecidas más adelante.

- Por ahora, estableceremos que un **mensaje** es una secuencia de símbolos, teniendo en cuenta que el orden de los símbolos es muy importante.
- La **principal función** de un mensaje es transportar información de un ente a otro.
- Los elementos de la comunicación deben estar de acuerdo en el mismo conjunto de símbolos, denominado **Alfabeto**.

Ejemplo 3

Denotamos por $\mathbb{A}$ el alfabeto que tiene 28 símbolos, las letras $\{A,B,C,D,E,F,G,H,I,J,K,L,M,N,\tilde{N},O,P,Q,R,S,T,U,V,W,X,Y,Z\}$ y un espacio, el cual es denotado por $\sqcup$ (28 ó 0).

Utilizaremos el alfabeto $\mathbb{A}$ para representar los mensajes escritos en Español.

Ignoramos elementos del idioma, distinción entre mayúsculas y minúscula, así como marcas de puntuación.

En un lugar de la Mancha ...

es reducido al siguiente mensaje utilizando el alfabeto $\mathbb{A}$.

$$\texttt{EN}\sqcup\texttt{UN}\sqcup\texttt{LUGAR}\sqcup\texttt{DE}\sqcup\texttt{LA}\sqcup\texttt{MANCHA}$$

Ejemplo 4

El alfabeto $\mathbb{B}$ tiene dos símbolos, 0 y 1, determinados dígitos binarios o bits.

Bits 0 y 1 son implementados electrónicamente como estados APAGADO y ENCENCIDO.

En la práctica, los bits son combinados en grupos, como palabras de 32 bits.

Cualquier mensaje transmitido electrónicamente es esencialmente una secuencia de bits.

Codificación

- **Codificar** es definido como una regla que permite una sustitución de un mensaje por otro mensaje.
- Los mensajes no tienen que compartir los mismos alfabetos.

Ejemplo 5

Una simple regla para codificar un mensaje en el alfabeto $\mathbb{A}$ *utilizando el mismo alfabeto es: escribir cada palabra al revés. Por lo que el mensaje:*

$$\texttt{NOS} \sqcup \texttt{VEMOS} \sqcup \texttt{PRONTO} \textit{ sería } \texttt{SON} \sqcup \texttt{SOMEV} \sqcup \texttt{OTNORP}$$

Ejemplo 6

Una regla para codificar mensajes en $\mathbb{A}$ *utilizando el alfabeto* $\mathbb{B}$ *es: reemplazar las vocales por ceros y las consonantes por unos. Por lo que el mensaje:*

$$\texttt{NOS}\sqcup\texttt{VEMOS}\sqcup\texttt{PRONTO}\ \textit{sería}\ 101101011110110$$

Hasta ahora ejemplos mostrados son artificiales tienen un valor limitado.

Debemos mirar los propósitos de la codificación para acercarnos a la realidad.

Hay tres razones principales para codificar un mensaje:

Económicas: Situaciones es necesario utilizar un alfabeto menor que los lenguajes naturales. O es deseable el mensaje más pequeño, esto ha dado lugar al desarrollo de técnicas para la *Compresión de datos*.

Fiabilidad: El proceso de transmisión no está libre de "ruido" que proporciona alteraciones en los mensajes. Necesario desarrollar *Corrección de los Errores*.

Seguridad: Usualmente es necesaria la confidencialidad de los mensajes. Históricamente, desarrollada en las comunicaciones diplomáticas y militares, pero hoy en día juegan en todas las transacciones económicas. Este área de codificación es conocida como *Criptografía*.

- Otros usos como *Algoritmos de Predicción* en el caso de dispositivos de tamaño reducido.

Ejercicio 1

Los siguientes mensajes son versiones codificadas del Español. Determine las reglas de codificación utilizadas y encuentre el mensaje original.

20 22 5 19 21 5

10100 10110 00101 10011 10101 00101

Definiciones Formales

Definición 1 (Alfabeto y Símbolos)

Un Alfabeto es un conjunto finito S. Los miembros de S son denominados símbolos.

Definición 2 (Mensaje, cadena y palabra)

Un mensaje en el alfabeto S es una secuencia finita de miembros de S:

$$x_1 x_2 \dots x_n (x_i \in S, 1 \leq i \leq n).$$

Un mensaje es una cadena de miembros de S, o una palabra en S, donde n es la longitud del mensaje, cadena o palabra.

El conjunto de todas las cadenas de longitud n es representado por S^n. Por ejemplo, cuando $S = \mathbb{B}$ y n=3, el conjunto de $\mathbb{B}^3$ está formado por las siguientes cadenas:

$$000\ 001\ 010\ 011\ 100\ 101\ 110\ 111$$

El conjunto de todas las cadenas en S es denotado como S^*:

$$S^* = S^0 \cup S^1 \cup S^2 \ldots S^{n-1} \cup S^n$$

Aunque S^0 consiste de la cadena con longitud cero, o la cadena sin símbolos.

Definición 3 (Palabra Codificada)

Sean dos alfabetos S y T y una función c inyectiva $c : S \to T^*$.
Para cada símbolo $s \in S$ la cadena $c(s) \in T^*$ es denominada la **palabra codificada** *para s.*

Definición 4 (Código)

El conjunto de todas las palabras codificadas, $C = \{c(s)|s \in S\}$, es denominada como **código**.
Cuando $|T| = 2$ el código es denominado binario, en el caso que $|T| = 3$ es ternario y en general cuando $|T| = b$, es b-nario.

Por ejemplo, sea $S = \{x, y, z\}, T = \mathbb{B}$, y definimos

$$c(x) = 0,\ c(y) = 10,\ c(z) = 11.$$

Tenemos un código binario, y el conjunto de palabras codificadas es $C = \{0, 10, 11\}$

- Un código c asigna a cada símbolo en S una cadena de símbolos en T, que pueden ser de distinta longitud.
  - Queremos construir un código para el alfabeto $\mathbb{A}$ utilizando el alfabeto binario $\mathbb{B}$.
  - Si establecemos palabras codificadas de longitud 4, sería como sigue:

$$A \mapsto 0000 \; B \mapsto 0001 \; C \mapsto 0010 \ldots$$

- c debe ser un función inyectiva, para que c no asigne el mismo código a dos símbolos diferentes. En otras palabras, si $c(s) = c(s')$ entonces $s = s'$.
  - En este caso solamente tenemos 16 cadenas de longitud 4 en $\mathbb{B}$, por lo que los 28 símbolos en $\mathbb{A}$ no pueden ser asignados a símbolos diferentes.

Previamente solamente hemos considerado la codificación de símbolos individuales, aunque puede realizarse una extensión a mensajes (cadenas de símbolos).

Definición 5 (Concatenación)

Si $c : S \to T^*$ es un código, extendemos c a S^* como sigue. Dado una cadena $x_1 x_2 \dots x_n$ en S^*, definimos:

$$c(x_1 x_2 \dots x_n) = c(x_1)c(x_2) \dots c(x_n)$$

Este proceso es conocido como concatenación, donde extendemos la función $c : S^* \to T^*$

No siempre es posible recuperar la cadena original desde la versión codificada. Por ejemplo, sea S = {x, y, z}, y define $c : S \to B^*$ como:

$$x \mapsto 0, y \mapsto 01, z \mapsto 10$$

Si tenemos la cadena codificada 010100 , que es el resultado de codificar una cadena en S^* utilizando c. En una primera aproximación podemos obtener dos resultados:

$$xzzx \mapsto 010100, yyxx \mapsto 010100$$

Esta situación debe ser evitada siempre que sea posible.

Definición 6 (Unívocamente Decodificable)

El código $c : S \to T^*$ es unívocamente decoficable (UD) si la función extendida $c : S^* \to T^*$ es una función inyectiva, por tanto cada cadena en T^* corresponde a un mensaje en S^*

A lo largo del curso, mostraremos como la propiedad UD puede ser garantizada.

Ejercicio 2

Un código binario es definido por las siguientes reglas:

$$s_1 \mapsto 10, s_2 \mapsto 010, s_3 \mapsto 100$$

Muestra mediante un ejemplo que el código no es unívocamente decodificable.

Código Morse

A	.-	M	--	Y	-.--	6	-....
B	-...	N	-.	Z	--..	7	--...
C	-.-.	O	---	Ä	.-.-	8	---..
D	-..	P	.--.	Ö	---.	9	----.
E	.	Q	--.-	Ü	..--	.	.-.-.-
F	..-.	R	.-.	Ch	----	,	--..--
G	--.	S	...	0	-----	?	..--..
H	....	T	-	1	.----	!	..__.
I	..	U	..-	2	..---	:	---...
J	.---	V	...-	3	...--	“	.-..-.
K	-.-	W	.--	4	....-	‘	.----.
L	.-..	X	-..-	5	.....	=	-...-

Código ASCII

DEC	HEX	CHAR	DEC	HEX	CHAR	DEC	HEX	CHAR	DEC	HEX	CHAR
0	00	NUL	32	20	SP	64	40	@	96	60	`
1	01	SOH	33	21	!	65	41	A	97	61	a
2	02	STX	34	22	"	66	42	B	98	62	b
3	03	ETX	35	23	#	67	43	C	99	63	c
4	04	EOT	36	24	$	68	44	D	100	64	d
5	05	ENQ	37	25	%	69	45	E	101	65	e
6	06	ACK	38	26	&	70	46	F	102	66	f
7	07	BEL	39	27	'	71	47	G	103	67	g
8	08	BS	40	28	(	72	48	H	104	68	h
9	09	HT	41	29	)	73	49	I	105	69	i
10	0A	LF	42	2A	*	74	4A	J	106	6A	j
11	0B	VT	43	2B	+	75	4B	K	107	6B	k
12	0C	FF	44	2C	,	76	4C	L	108	6C	l
13	0D	CR	45	2D	-	77	4D	M	109	6D	m
14	0E	SO	46	2E	.	78	4E	N	110	6E	n
15	0F	SI	47	2F	/	79	4F	O	111	6F	o
16	10	DLE	48	30	0	80	50	P	112	70	p
17	11	DC1	49	31	1	81	51	Q	113	71	q
18	12	DC2	50	32	2	82	52	R	114	72	r
19	13	DC3	51	33	3	83	53	S	115	73	s
20	14	DC4	52	34	4	84	54	T	116	74	t
21	15	NAK	53	35	5	85	55	U	117	75	u
22	16	SYN	54	36	6	86	56	V	118	76	v
23	17	ETB	55	37	7	87	57	W	119	77	w
24	18	CAN	56	38	8	88	58	X	120	78	x
25	19	EM	57	39	9	89	59	Y	121	79	y
26	1A	SUB	58	3A	:	90	5A	Z	122	7A	z
27	1B	ESC	59	3B	;	91	5B	[	123	7B	{
28	1C	FS	60	3C	<	92	5C	\	124	7C	|
29	1D	GS	61	3D	=	93	5D	]	125	7D	}
30	1E	RS	62	3E	>	94	5E	^	126	7E	~
31	1F	US	63	3F	?	95	5F	_	127	7F	DEL

Codificación PEM

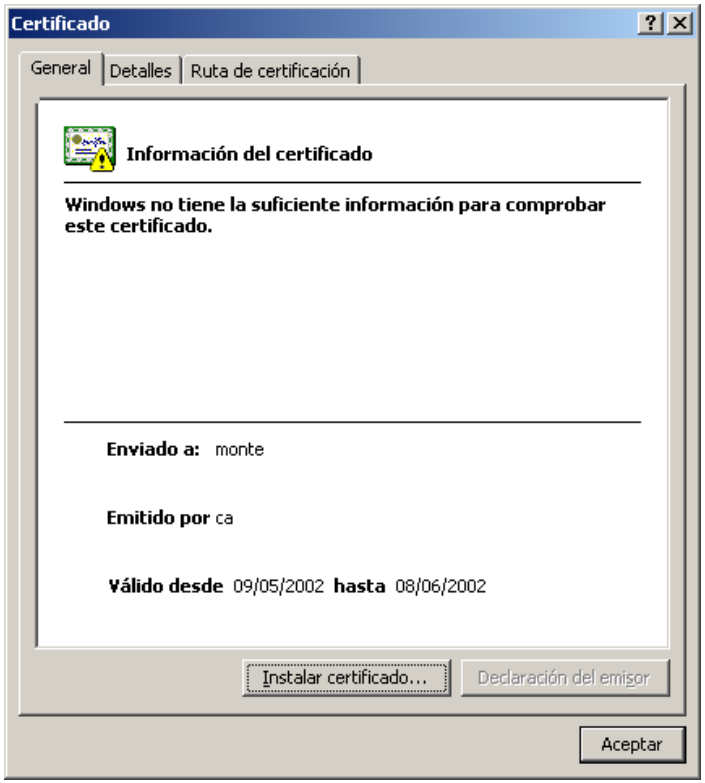

Figura 1 : Certificado Clave Pública. Windows

```
-----BEGIN CERTIFICATE -----
DwYDVQQLEwhEcHRvIExDQzEOMAwGA1UEAxMFbW9udGUxHzAdBgkqhkiG9w0BCQEW
EG1vbnRlQGxjYy51bWEuZXMwgZ8wDQYJKoZIhvcNAQEBBQADgY0AMIGJAoGBALJS
Ms+RrNOlt55b5peRaLvNjwiTCbWAZDBhMeL23qx44mLbXRsf2Ji/kkVmK3psksY7
jekYaDntGVpvBydoWBsnMgmD/Ik5yQ7iteAuFwT3fprHoG3oJ3RrIvXT7zTwDlrv
MT9fPxx6BpIPBYKWK0cMPLhgJfHwxMfPNYIb3bvhAgMBAAGgADANBgkqhkiG9w0B
AQQFAAOBgQBp2AMVHtow2SRqRenemrmyKzH8xYlothzPKcqSgC34bDec5uzIMJYw
WKILtI0TqZnZvNsKZEeDhqr6whWcTGdGCI9SaeZCKy02BfMrSR+WiwndOq8Rrz7J
6QXEBLdeVr4sR7WlwNSKZAxiVHydS8MLbuRhZV3eQm7yEi0UJe22qA==
-----END CERTIFICATE -----
```

Figura 2 : Certificado Clave Pública. PEM

Codificación de datos 2D

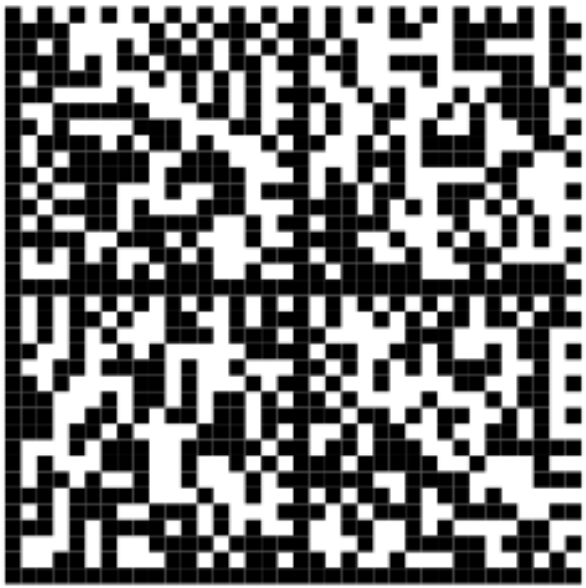

Figura 3 : Datamatrix

Figura 4 : QRCode

Ejemplo Doble Codificación

Figura 5 : Billete de IDA

72330001599105441351003260211139030010330169 5DUHK

7233000159910 5441351003 **260211 13903 001 033** 016 **95DUHK**

Num. Billete: 7233000159910

Fecha: 260211

MD: 13903

Coche: 001

Plaza: 033

Localizador: 95DUHK

Figura 6 : Billete de Vuelta

7233000159910 5441351003 **260211 13903 001 033** 016 **95DUHK**
7233000159936 5100354413 **280211 13904 001 030** 018 **95DUHK**

José A. Montenegro Montes
Dpto. Lenguajes y Ciencias de la Computación
ETSI Informática. Universidad de Málaga
monte@lcc.uma.es
twitter

UNIVERSIDAD DE MÁLAGA

E.T.S. INGENIERÍA INFORMÁTICA

LENGUAJES Y CIENCIAS DE LA COMPUTACIÓN
UNIVERSIDAD DE MÁLAGA