IV: Estructura y metodología de la codificación

Por Alejandro Coreltti (acorletti@hotmail.com)

1. Introducción:

El ser humano recibe información del mundo exterior a través de sus sentidos, ingresa esta información bajo la forma de imágenes, sonidos olores, temperaturas, etc. Esta información, es procesada y almacenada en el cerebro. Al tener que transmitirla es donde se plantea el problema que si bien hoy parece simple, también llevó muchas décadas a la humanidad, y sigue avanzando a través de las nuevas técnicas de comunicación que aparecen día a día. Si se hace un análisis cronológico, se nota claramente que en los albores de la humanidad, el conocimiento se transmitía de maestro a discípulo o de artesano a aprendiz en forma verbal, todo el conocimiento se perdía al fallecer una persona pues este no quedaba plasmado en ningún medio. Al aparecer las primeras formas de escritura, el conocimiento queda asentado y se transmite más allá del conocimiento personal.
Como se puede apreciar en el párrafo anterior, se plantean dos problemas:

- Los *símbolo*s empleados para transmitir la información.
- Los *medios empleados para almacenarla* (piedra, papel, tela, etc).

Cualquier lector de este texto puede interpretarlo en virtud de conocer el idioma con que está escrito, pero no es ajeno a la existencia de muchos otros lenguajes, los cuales pueden o no basarse en los mismos símbolos, pues por ejemplo el Idioma Inglés se basa en casi el mismo Alfabeto, pero no igual pues por ejemplo no existe la "ñ" y existe "&"; también si analizamos el idioma Chino, este no emplea ni siquiera los mismos símbolos, como tampoco su sintaxis.

> En resumen, existen distintos **Idiomas o Lenguajes**, estos pueden emplear el mismo o diferentes conjuntos de **símbolos o códigos**, el conjunto de códigos de un idioma respecto a otro no necesariamente tiene que tener la misma cantidad de símbolos

En la actualidad existe una gran variedad de medios de almacenamiento de información, como son el papel, las cintas, los diskette, los CD, los videos, etc.

2. Codificación de la Información:

Cuando la información que originariamente se encontraba representada en un Alfabeto A1, se convierte a un segundo Alfabeto A2, se dice que ha sido Codificada.

El caso más sencillo es cuando existe una correspondencia *biunívoca o biyectiva* entre ambos alfabetos, es decir a cada símbolo de A1 le corresponde uno y sólo uno de A2 y viceversa. Esta condición no siempre se cumple. El caso más conocido es el del Código Morse, sumamente difundido pues es quien da origen a la transmisión de información a distancia a través del Telégrafo. Si se analiza el funcionamiento de este sistema de codificación, se distingue claramente el concepto de **Alfabeto y Codificación**, pues este sistema solo dispone de 2 códigos o símbolos: *Punto y Raya* (es decir es un sistema binario = solo dos estados). Por lo tanto si se desea transmitir un texto en Castellano que posee 26 símbolos a través de un telégrafo que sólo tiene dos símbolos, no existirá una correspondencia biunívoca entre ambos, es por esta razón que se debe hacer empleo del concepto de codificación, asociando a cada símbolo del alfabeto Castellano un conjunto de

símbolos del Alfabeto Morse, codificando de esta forma la información (Ej: **A = .- , I = .. , T = - , Z = --..** , etc) . Otra característica que posee este código es la de no ser de longitud fija, cualidad que se tratará más adelante, pero que no se desea pasar por alto. También se verá más adelante que el pasaje a información binaria es lo cotidiano para dialogar con un ordenador.

Existen también códigos que poseen la singularidad que a un mismo carácter destino es decir una misma secuencia binaria, pueden corresponder más de un símbolo origen, es decir más de una letra en Alfabeto Castellano. Aunque parezca extraño estos códigos existen y justamente se llaman *Códigos singulares*. El mejor ejemplo de estos es el Código Baudot o Alfabeto Internacional Nro 2 que se tratará más adelante.

NECESIDAD DE CODIFICAR LA INFORMACION:

- Transmisión automática de Información.
- Abreviar la escritura (Matrículas, Formularios, Códigos de artículos, etc).
- Hacer secreta o ininteligible la Información (Cifrado).

3. Códigos de longitud fija y variable:

Hasta ahora no se ha hecho referencia, pero queda claro que el código Morse puesto de ejemplo no posee la misma longitud de codificación en cada carácter, este en particular fue diseñado así adrede, para aprovechar la probabilidad de ocurrencia de los distintos caracteres, asignándole una menor cantidad de símbolos a los que más ocurrencias de aparición poseen. Estos códigos se llaman de *longitud variable o Código Compacto* y se estudian a fondo a través de la Teoría de la Codificación, cuyos precursores fueron Shannon y Hartley.

En la actualidad por el contrario, la masa de los protocolos de comunicaciones y Sistemas operativos, suelen emplear Códigos de *longitud fija o Códigos Bloque* es decir que cada uno de sus caracteres es representado por igual cantidad de símbolos, en general estos códigos facilitan la velocidad de codificación y decodificación, optimizando el rendimiento del sistema, como se verá a continuación, los ordenadores actuales, emplean conjuntos definidos de información que ya casi están Universalmente estandarizados.

4. Sistema de codificación binario:

4.1. Introducción:

En el capítulo anterior se mencionó la transmisión analógica y la digital. Hoy en día la transmisión digital es la protagonista absoluta de la escena. Esta situación se basa en la capacidad que tiene para ser **Regenerada** infinitas veces, siendo cada vez idéntica a la señal original actividad que no se puede realizar con la transmisión analógica pues esta al ser **Amplificada**, cada vez que se lo hace incorpora ruido o distorsión que se va sumando a la señal original, llegando un momento en el que es ininterpretable. El pasaje del mundo digital al analógico y viceversa, hoy se puede realizar en todo tipo de señal. Este fenómeno es llamativo pues actividades netamente analógicas como son la voz, la música o el vídeo, hoy son de mucha mayor calidad si se las trata digitalmente (CD, DVD, telefonía celular, etc).

El concepto de la transmisión binaria va asociado al pasaje o no de corriente o luz al cierre o apertura de una compuerta o relé, etc, de forma tal que se puede asociar un valor de 1 (uno) o 0 (cero), acorde a la presencia o no de energía o señal.
Un elemento biestable en el que se pueden diferenciar netamente dos estados se llamará Variable binaria.
Como el término Dígito Binario es sumamente extenso, se convino en abreviarlo y llamarlo **BIT** por **BI**nary Digi**T**).

4.2. Parámetros considerados en la construcción de un código:

a. **Eficacia**: El número de dígitos binarios no podrá ser superior a Log_2 N, siendo N el Número de símbolos a codificar. Cuanto más cercano esté el resultado de este logaritmo a la cantidad de dígitos binarios empleados mayor será la eficiencia, siendo del 100 % si los valores son iguales.

b. **Valores numéricos**: El código deberá ser capaz de reconocer cifras decimales para su posterior tratamiento matemático.

c. **Ordenamiento**: EL código debe facilitar las clasificaciones u ordenamientos habituales, como por ejemplo el ordenamiento alfabético.

d. **No singularidad**: No repetición de símbolos.

e. **Unívocamente decodificable**: Que no se preste a ambigüedades en su extensión, es decir que al concatenar los símbolos no se pueda presentar la posibilidad de poder ser interpretado con distintas opciones.

f. **Decodificación instantánea**: La presentan los códigos que incorporan la capacidad de prefijo. Al llegar determinado símbolo o concatenación de estos, automáticamente se sabe que ha finalizado ese carácter.

g. **Seguridad**: Es deseable que el código proteja la información respecto a la transmisión, facilitando la detección y/o corrección de errores.

5. <u>Medidas de Información:</u>

5.1. BIT: estado lógico equivalente a 1 o 0.

5.2. Byte u Octeto: Agrupación de 8 bit.

Esta definición es la que realmente se universalizó para el tratamiento de la información y la palabra octeto es hoy una de las bases de la transmisión de información, la razón de ser de esta convención radica en:

- La capacidad suficiente de codificación que posee un octeto es decir 256 posibilidades diferentes.

Si se plantea el conjunto de posibilidades este irá desde: 0000 0000 , 0000 0001 , 0000 0010 , 0000 0011 , 0000 01000.....1111 1111.
Ante lo cual permite hasta 256 códigos diferentes.

- El fácil pasaje entre el sistema decimal, hexadecimal y binario.

Suma Decimal	128 + 64 + 32 + 16 + 8 + 4 + 2 + 1 = 256								
Peso Decimal	128	64	32	16	8	4	2	1	
Binario	b	b	B	b	b	b	b	b	
Peso hexadecimal	8	4	2	1	8	4	2	1	
Suma hexadecimal	8 + 4 + 2 + 1 = F				8 + 4 + 2 + 1 = F				FF
EJEMPLO									
Suma Decimal	128 + 32 + 2 + 1 = 163								
Peso Decimal	128	0	32	0	0	0	2	1	
Binario	1	0	1	0	0	0	1	1	
Peso hexadecimal	8	0	2	0	0	0	2	1	
Suma hexadecimal	8 + 2 = A				2 + 1 = 3				A3

Hexadecimal: Conjunto de16 símbolos (0, 1, 2, 3, 4 ,5 , 6, 7, 8, 9, A, B, C, D, E ,F).

5.3. CARÁCTER: Es la unidad de información a nivel alfabeto humano, representa cualquier símbolo del alfabeto usado como alfabeto normal. Se los clasificará en:

a. **Alfabéticos**: Letras en mayúsculas y minúsculas.

b. **Numéricos**: dígitos de 0 a 9.

c. **Especiales**: Puntuación, paréntesis, operaciones aritméticas y lógicas, comerciales, etc.

d. **De operación y control**: Destinados al control de la transmisión de Información (Retorno de carro, nulo, SYN, ACK, DLE, EOT, SOH, etc)

5.4. Bloque, Mensaje, Paquete, Trama: Son distintas formas de agrupamiento de Byte, y se definen acorde a las distintas técnicas de transmisión de información o Protocolos de Comunicaciones.

6. Tipos de codificación binaria:

6.1. Código Hollerith:

Este sistema de codificación diseñado por Hermann Hollerith, toma por ejemplo las tarjetas perforadas que controlaban telares, inventadas por Jacquard. Se basa en tarjetas de 80 columnas por 12 filas de perforaciones, siendo cada fila la representación de un carácter, es decir que por cada tarjeta se podían representar hasta 80 caracteres. Como cada columna tiene doce posibles orificios, se trata de un código de 12 bit.

Si bien las tarjetas perforadas duraron mucho tiempo, en la actualidad es un medio que ya no existe, razón por la cual no se profundizará más sobre este tema.

6.2. Código para el servicio de Telex (Alfabeto Internacional Nro 2) (Código Baudot):

Este código lo diseñó Baudot en el Siglo XIX (Quien dio origen a la Unidad de medida de la velocidad de modulación [Baudio]), y consta de 5 bit, por lo que si bien cabe esperar que pueda codificar 32 símbolos, en realidad sirve para codificar 58. Para poder implementar esta técnica de codificación SINGULAR, posee dos caracteres especiales (11111 = letras , 11011 = dígitos). Al accionar una de estas teclas, automáticamente todos los caracteres que siguen a continuación serán interpretados acorde al significado de esta tecla (por Ej: letras), si luego se desea cambiar al otro conjunto, se deberá presionar la tecla contraria (por Ej: dígitos) y a continuación todos los caracteres que siguen se interpretarán como dígitos.

6.3. Código BCD (Binary Coded Decimal – Decimal Codificado en Binario):

Si se emplean cuatro bit, es posible obtener 16 posibles combinaciones (2x2x2x2= 16, es decir 2^{16}), como se mencionó anteriormente basado en la notación posicional, cada bit tiene un peso acorde a la posición que ocupe en el cuarteto (8, 4, 2, 1). Si se desea realizar un pasaje directo desde este código binario de longitud fija (4 bit) a un sistema decimal, lo más rápido es realizar la operación en forma directa sumando su peso relativo.

Ej:

Peso 8	Peso 4	Peso 2	Peso1	Decimal
0	0	0	0	**0**
0	0	0	1	**1**
0	0	1	0	**2**
0	0	1	1	**3**
0	1	0	0	**4**
0	1	0	1	**5**
0	1	1	0	**6**
0	1	1	1	**7**
1	0	0	0	**8**
1	0	0	1	**9**

Si bien se desperdician 6 posibilidades, se beneficia con la velocidad de codificación.

6.4. Código EBCDIC (Extended Binary Coded Decimal Information Communication):

Este código fue diseñado por IBM, la idea básica es anteponer al código BCD 4 posiciones más, con lo que se obtiene un código binario de longitud fija de 8 bit. Fue y sigue siendo muy usado internamente en los ordenadores, y se lo suele ver representado a través de pares de números hexadecimales. EL lenguaje de programación ASSEMBLER, pionero de la programación se realizaba (Gracias a Dios ya no) sobre la base de este Código, razón por la cual era altamente eficiente (e inhumano), pues se programaba directamente en lenguaje de máquina.

6.5. Código ASCII (American Standard for Computers Information Interchange):

Es un código de 7 bit que ha llegado a ser el estándar mundial de transmisión de Información.
Este se impuso luego de la Conferencia Plenaria que se celebró la ITU (International Telecommunication Union) en el año 1968 en la Ciudad de Mar del Plata (Argentina).

6.6. Alfabeto Internacional Nro 5:

Este código es muy similar al ASCII, pero contempla particularidades de cada País y otra especiales referidas a los idiomas diferentes al inglés.
Posee una Versión Internacional de referencia (VIR) que proporciona la flexibilidad anteriormente mencionada. Se logra estandarizar a través de dos grandes organizaciones como son ITU e ISO (International Standard Organization).

6.7. Código PC-8:

Es el código que realmente se usa en las computadoras personales, y como está basado en el ASCII, muchas veces se lo confunde con este. Este código es de 8 bit y es el que bajo sistemas operativos DOS o Windows aparece al presionar la tecla [ALT]. A partir de este código es que se da origen a todos los conjuntos de caracteres que ofrece hoy cualquier procesador de texto (New Roman, Sans serif, arial, etc). Ese código, toma como base la mencionada VIR, es por esto que muchas veces se lo sule llamar ASCII ampliado.

6.8. Código 4 de 8:

Aunque es un código de 8 bit, fue diseñado como uno de 6 bit en cuanto a las posibilidades de codificación que este ofrece y para poder convertirlo fácilmente a BCD. Las únicas combinaciones válidas son aquellas dónde **4 de los 8** bit son ceros y los otros 4 son unos, así por ejemplo la combinación 1100 0100, 0100 1000, 0000 0001, 1111 1100, no son válidas por no poseer cuatro 1 y cuatro 0. La totalidad de los códigos que ofrece son 70, y como BCD sólo tiene 64, hay 6 de ellos que no pueden traducirse, los cuales son funciones de control que BCD no usa. La redundancia que impone este código, es empleada para la detección de errores, y el precio que se paga es que de 256 combinaciones, solo emplea 70.

7. Verificación de códigos:

A pesar de contar con sistemas de transmisión altamente confiables, estos no son perfectos y la posibilidad de alteración del estado de uno o varios bit no es nula, razón por la cual, muchos de los códigos mencionados emplean medidas que permiten verificar la integridad de la información.

La técnica más simple es el control de paridad que consta de agregar bit adicionales, los cuales a través de la suma binaria del símbolo transmitido, se obtendrá un valor 0 o 1. Acorde a este resultado, se puede tomar como convención agregar un bit adicional que identifique claramente el resultado de esta suma, esta convención se puede optar a través de una metodología PAR o IMPAR, es decir que al integrar este nuevo bit se determinará la paridad o imparidad del resultado. Si se opta por el control PAR, la suma de la totalidad de los bit (incluyendo el de paridad) deberá dar como resultado siempre un número par, ante lo cual, si la suma del carácter sin paridad daba PAR, el bit de paridad a incluir será un 0 y por el contrario si la suma del carácter sin paridad daba IMPAR el bit de paridad a incluir será un 1; como se puede apreciar en ambos casos la suma total de los bit (datos y bit de paridad) dará como resultado SIEMPRE PAR. Si se decidiera el empleo de paridad IMPAR, se implementa de modo inverso, obteniendo como resultado SIEMPRE IMPAR en la suma de la totalidad de los bit (datos y bit de paridad).

El equipo que recibe esta información deberá estar configurado con la misma lógica que el emisor es decir ambos PAR o ambos IMPAR, pero no se puede configurar distinto. Al ir recibiendo los

distintos caracteres, irá realizando la misma suma que hizo el emisor, luego de la cual comparará el bit de paridad, el cual si coincide está correcto, pero si en su suma detecta que debía ser PAR y encuentra que el resultado es IMPAR (o viceversa), sabrá fehacientemente que existió un error a lo largo del sistema.

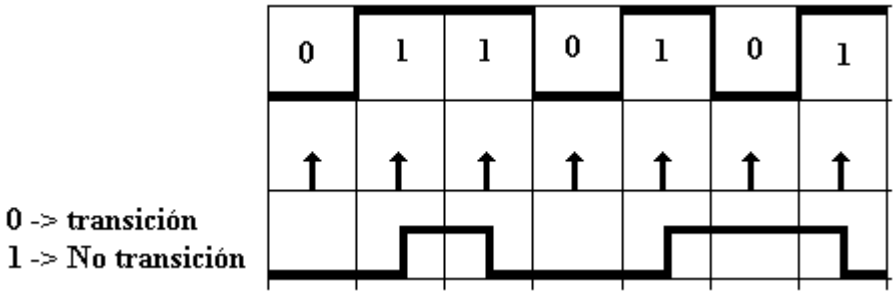
Es claro también que si se alteran 2 bit o cualquier número par de estos, la suma de verificación coincidirá perfectamente, no detectándose el error. Es por esta razón que esta técnica es de las más básicas que se emplean, existiendo en la actualidad estrategias mucho más complejas para esta actividad.

8. Tipos de codificación en banda base:

Al transmitir información en banda base, esta será unipolar, polar, o bipolar con o sin retorno a cero; pero a su vez, existen diferentes metodologías de generar las transiciones entre un estado de tensión y otro. Es en realidad la lógica que se emplea para la transmisión física de información. Cada una de estas lógicas se diseñan para sincronizar la transmisión, minimizar o maximizar las componentes de continua (secuencias seguidas de unos o ceros), aumentar la velocidad de transmisión respecto a la de modulación, etc. Estas técnicas de codificación se detallan a continuación.

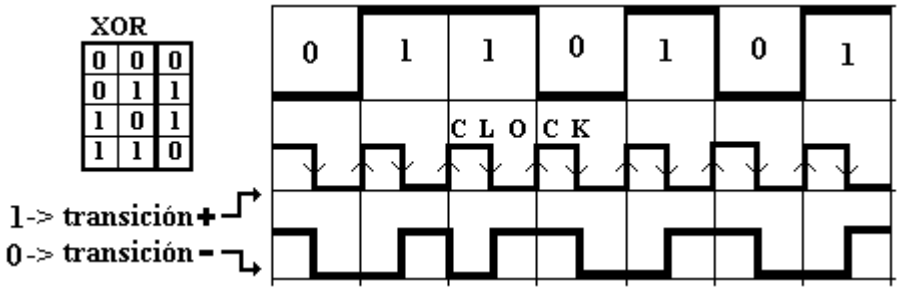
8.1. Diferencial:

Al presentarse un uno existirá una transición, ante un cero no.



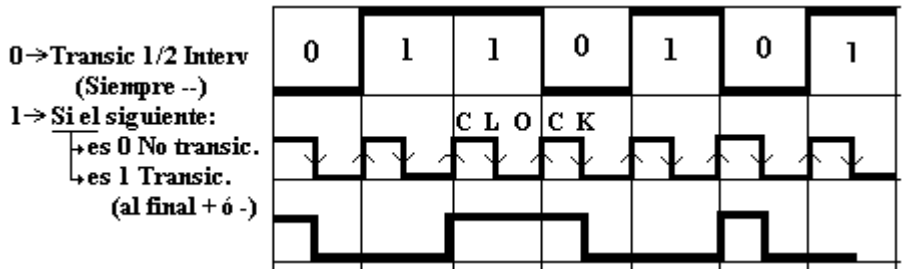
8.2. Manchester:

Ante la presencia de un uno, existirá una transición positiva ante un cero será transición negativa. Con esta técnica se elimina la componente de continua pues siempre existirá una transición, es por esta razón que se trata de un código autosincronizante pues al ir recibiendo cada bit, se sabrá perfectamente dónde empieza y termina, pues existirá una transición. El inconveniente que posee es que duplicará el ancho de banda empleado. Como se representa a continuación, también se puede pensar su regla de formación a través del XOR (or exlusivo).



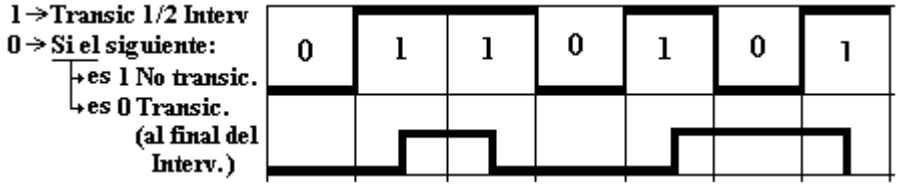
8.3. Manchester diferencial:

Ante la presencia de un cero, existirá una transición siempre negativa en la mitad del intervalo. Ante la presencia de un cero, si el siguiente es cero, no habrá transición, si es un uno, existirá transición positiva o negativa al final del intervalo.



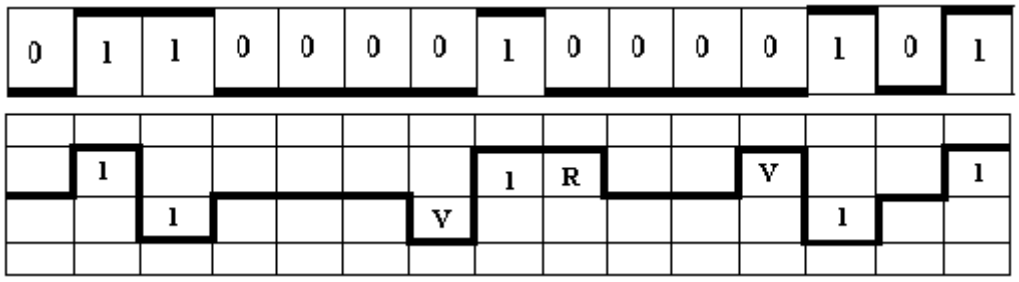
8.4. Miller:

Ante la presencia de un uno existirá transición en la mitad del intervalo. Ante un cero si el siguiente es un cero habrá transición al final del intervalo, si es un uno no habrá transición. Este método garantiza por lo menos una transición cada dos intervalos significativos, requiriendo con esto menor ancho de banda que el código Manchester.



8.5. HDB – 3:

Se basa en el código AMI (Alternative Mak Inversion), que es en realidad el Bipolar sin retorno a cero, pero a este le agrega una novedad que permite eliminar las componentes de continua producidas por secuencias de ceros. Esta técnica agrega un bit llamado bit de violación, el cual se colocará siempre que se presente una secuencia de tres ceros seguidos, y se distinguirá por encontrarse con la misma polaridad que el último uno aparecido (es decir sin inversión). Para reducir la posibilidad de errores, propone también dos tipos de bit de violación, aquel que se presenta luego de una secuencia par de unos desde la última violación y el que se colocará luego de una secuencia impar de unos desde la última violación, como se detalla a continuación.



8.6. 4B – 3T:

Este código toma secuencias de 4 bit binarios (dos niveles) y los convierte en ternas de bit de tres umbrales de detección, reduciendo a un 75 % el ancho de banda. En esta codificación se desperdicia capacidad de codificación pues sobre $2^4 = 16$ códigos fuente se transforman a un espacio posible de $3^3 = 27$, dejando de lado entonces 27 – 16 = 11 códigos no utilizados.

Binario	**Ternario**
0000	0 –1 +1
0001	-1 +1 0
0010	-1 0 +1
.....	.....
.....	.....
1111	0 +1 -1