第一章

編 碼 理 論 與 技 術 簡 介

課程內容

- **1.1** 信息理論介紹
- **1.2** 信息源
- **1.3** 常用編碼

1.1信息理論介紹

- 信息理論(Information Theory)是由沈農(Shannon)奠基的一門嶄新的數學學科，它適用於有效而可靠的通訊通道中。信息理論的基本問題是研究有效而可靠地傳遞信息的可能性與方法。
- 1948 年，美國工程師C.E.Shannon 在貝爾實驗室出版的專門期刊上，發表了“通訊的數學理論”，在這一開創性文章中，給出了信息度量的數學公式，也為信息理論的創立做出了獨特的貢獻。

$$C = B\log_2(1+SNR)$$

- 由這個公式，開創通訊理論的新領域—信息理論，並對後來發展出的編碼理論(Coding Theory)及密碼學(Cryptology)產生很大的影響。我們可以利用信息理論做為工具，從理論上定量分析密碼系統的安全，也可以計算出通訊頻道的容量和限制。

- 沈農定義的信息量，撇開了事件發生的時間、地點、內容以及事件與人們的情感及人們對事件的反應，而只顧及事件發生的狀態數目及每種狀態發生的可能性大小，這就信息度量具有普遍意義和廣泛的適應性。
- 為了確保證信息傳遞的可靠性，必須將所傳遞的信息進行編碼並對接收到的信息進行解碼，因此，信息理論的基本問題既是在給定信號源及傳遞信息通道的條件下，選擇適當的編碼與解碼，使接收信息與發送信息不一致的機率儘可能的小。

- 信息理論為在可能的信息材料範圍內選取信息，並且將它們製作成任何形式可在通道傳送的信號。
- 沈農以極嚴謹的理論證明0 與1 的資料壓縮與傳輸，實際上存在壓縮極限與傳輸極限，前者為資料的熵(Entropy)，後者為傳輸通道容量 (Channel Capacity)。

- 信息理論的基本理論、編譯碼技術、密碼學構成了信息理論的三個重要的相互聯繫的組成部分。
- Shannon 的理論廣適於類比信號與離散信號源兩者，因為數位化系統日趨普遍之故，大都討論離散部份。
- 沈農用機率的方法來計算信息量，發生的機率越小，信息量越大，一件事情如果我們早已知道，再講出來，就是一點信息量也沒有。

- 例如：炎熱的七月天，氣象局突然宣佈要下雨，這對我們產生很大的衝擊，也就是說，我們獲得了很大的信息量，如果是明天，太陽會從東方升起，這信息讓我們覺得很普通。

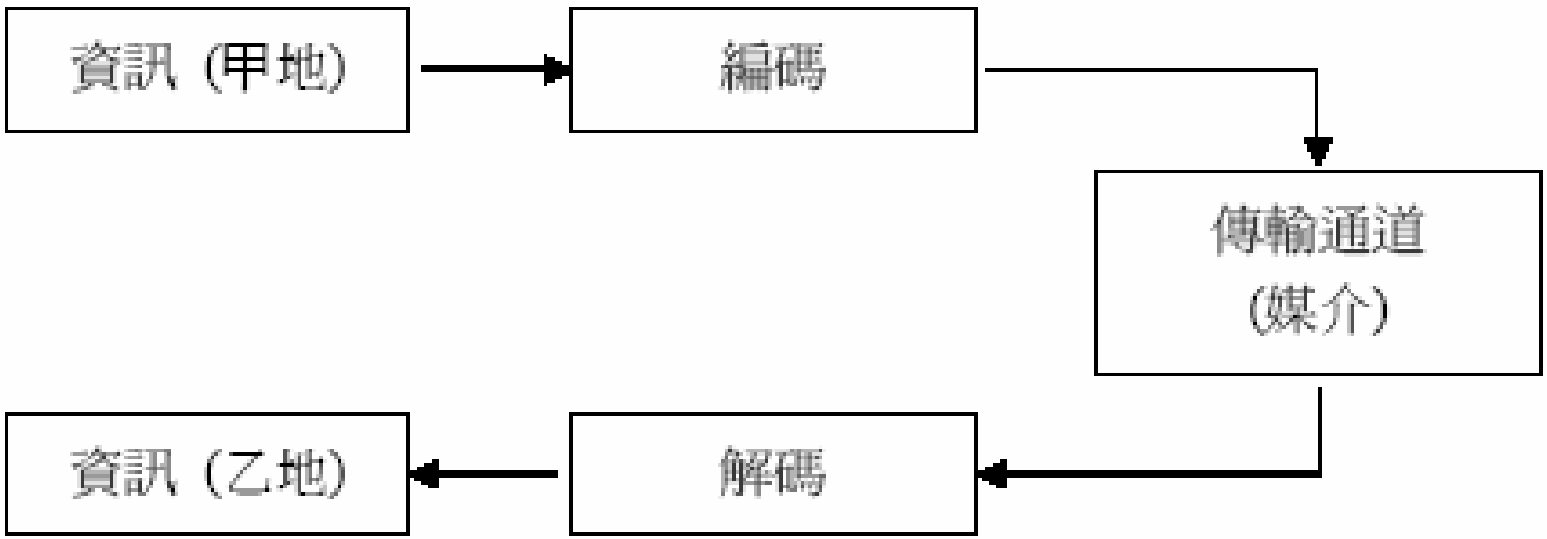

圖 1.1-1　資訊處理系統

1.2信息源

- 信息源(或稱為信號源，或簡稱為信源)(information source)就是一個機率場，例如，全體中文及其機率分佈（每個字的使用率)；26 個英文字母及其機率分佈等等，都是信息源。
- 通訊環境中最主要的信息源有下列四種：語音、影像、文字以及數據。
- 由字符符號的集合定義了信息源，這個符號集合稱之為信息字母集合。
- 例如：信息字母集合包括M 符號，其可表示如下

$$S = \{s_0, s_1, s_2, \cdots s_{M-1}\}$$

- 信息符號 $s_i \in S$ 的發生機率可以寫成為

$$p_i = p(s_i)$$

- 因為所有發生機率的總合等於1，即

$$\sum_{i=0}^{M-1} p_i = 1$$

- Shannon定義每個信息符號傳送的平均信息數量的測量。這個測量稱作信息端的熵(*Entropy*) 。其定義為

$$H(S) = \sum_{i=0}^{M-1} p_i \log(1/\, p_i)$$

- 如果我們以2 為對數值的基底去，熵測量的單位就叫作“位元”。
- 若是我們取用自然對數，熵測量的單位則稱作“自然單元(或是nat)”

- 在第二章，對熵會有詳細的介紹。
- **例題1.2-1:** 一個M=4 的信息端具有可能符號 s_1，s_2，s_3 與 s_4 用以表示一對二位元數字。符號 s_1 代表00，其發生機率為0.15；符號 s_2代表01，其發生機率為0.05；s_3 表示10，其發生機率0.3；s_4 表示11，其發生機率為0.5。請找出平均編碼長度和熵之值。

<解答> 熵 $= H(S) = \sum_{i=1}^{4} p_i \log_2(1/p_i) = 1.6447$ 位元

平均編碼長度 = 2位元

- 在例題1.2-1 中，信源端的四種符號的二位元資料描述需要兩個位元。
- 信源端熵的位元數小於平均編碼長度時，使用較少位元數設計出有效率的編碼機制是有可能的。
- 這稱為資料壓縮(*data compression)*，而用作這種目的的編碼器就是信源編碼器。
- 使用上述編碼機制便稱為信源編碼(*source coding)* ，其所產生的碼稱為信源碼(*source code)* 。

1.3 常用編碼

- 電腦系統中資料儲存的最基本單位是「位元(bit)」，而每個「位元」所能代表的數值只有 0 與 1 而已。
- 因此當我們要用電腦做資料的儲存與處理時，我們通常會將連續幾個位元和起來看成一個單位，稱之為「位元組 (byte)」，其所能代表的數值就不會只限於 0 與 1，而可以達到$2^n - 1$ (其中n 為該位元組中所包含的位元個數) 。
- 如此我們就可以將我們想記錄在電腦系統的符號一一編號，以位元組為單位儲存在電腦中。

- 編碼在數位通信裡，佔有極重要的角色，也就是說數位通信，可以偵錯、改錯、壓縮、保密。全是編碼的功能。
- 一個二進制可以表示兩種狀態，如 0 和 1。兩個二進制則可以描述四種相異狀態，如 00, 01, 10 和 11。
- 電腦系統是發源於美國，最早的編碼系統包括26個英文字母 (包括大小寫)、標點與其他特殊符號、外加一些控制碼。因此只需 7 個位元即包含所有所需的信息，總共可容納 128 個符號，這也就是大家所熟知的 ASCII編碼。

- 對於歐洲語系而言，7 個位元的編碼空間不符合所需，因為除了 26 個英文字母外，歐洲許多國家還需要拉丁字母、特殊字母的上下標與其他符號等，才能完整表示他們的語言。
- 因此原來的編碼系統必須再擴充，由 7 個位元變為 8個位元，以容納那些多出的符號，這就是 ISO8859 的編碼標準。

- 亞洲地區，上述的 8 位元編碼又不夠用了。因為亞洲地區不再是拼音文字，因此多出了成千上萬個符號，我們已無法仿照當初 ISO8859 的做法，
- 唯一的作法只有合併幾個位元來共同表示一個符號，如此才有足夠的空間來容納一個地區所有的文字。例如我們台灣地區最常見的 BIG5 編碼，就是將兩個位元組合起來看成一個字。

BCD 碼

- BCD 碼是由英文(Binary coded Decimal codes)而來的，稱2 進制10 進數碼。它是以四個位元的前10 種組合來表示10 進制中從0~9 共10 個數字。
- 例如，0000 → 0, 0001 → 1, 0010 →2,

  0011 → 3, 0100 → 4, 0101→5, 0110→6, 0111→7, 0001→8, 1001 → 9

ASCII 碼

- ASCII 碼是(American Standard for Information Interchange 簡寫ASCII，唸ASKEE)的縮寫而來，稱為美國標準信息交換碼，它是由美國的信息及電腦界所制定的信息交換標準。ASCII 碼最早，共使用7 位元二進位碼來編組，故又稱
- 為ASCII-7 碼，其共使用3 個區域位元，4 個數字位元。在ASCII Code 的表示法中，數字愈大其碼也相對愈大，英文字母大小寫也是愈後面的字母其碼也相對愈大。
- 例英文字母「A」的10 進制ASCII 值為65，則字母「Q」的10 進制ASCII值為81。

表 1.4-1 ASCII Code

Dec	Char	Dec	Char	Dec	Char	Dec	Char	Dec	Char	Dec	Char	Dec	Char	Dec	Char
0	NUL(null)	16	DLE(data-link-escape)	32	SPACE	48	0	64	@	80	P	96	`	112	p
1	SOH(start-of-heading)	17	DC1(device-control	33	!	49	1	65	A	81	Q	97	a	113	q
2	STX(start-of-text)	18	DC2(device-control	34	"	50	2	66	B	82	R	98	b	114	r
3	ETX(end-of-text)	19	DC3(device-control	35	#	51	3	67	C	83	S	99	c	115	s
4	EOT(end-of-transmission)	20	DC4(device-control	36	$	52	4	68	D	84	T	100	d	116	t
5	ENQ(enquiry)	21	NAK(negative-acknowledge)	37	%	53	5	69	E	85	U	101	e	117	u
6	ACK(acknowledge)	22	SYN(synchronous-idle)	38	&	54	6	70	F	86	V	102	f	118	v
7	BEL(bell)	23	ETB(end-of-trans.-block)	39	'	55	7	71	G	87	W	103	g	119	w
8	BS(backspace)	24	CAN(cancel)	40	(	56	8	72	H	88	X	104	h	120	x
9	TAB(horizontal-tab)	25	EM(end-of-medium)	41	)	57	9	73	I	89	Y	105	i	121	y
10	LF(NL-line-feed,-new-line)	26	SUB(substitute)	42	*	58	:	74	J	90	Z	106	j	122	z
11	VT(vertical-tab)	27	ESC(escape)	43	+	59	;	75	K	91	[	107	k	123	{
12	FF(NP-form-feed,new-page)	28	FS(file-separator)	44	,	60	<	76	L	92	\	108	l	124	|
13	CR(carriage-return)	29	GS(group-separator)	45	-	61	=	77	M	93	]	109	m	125	}
14	SO(shift-out)	30	RS(record-separator)	46	.	62	>	78	N	94	^	110	n	126	~
15	SI(shift-in)	31	US(unit-separator)	47	/	63	?	79	O	95	_	111	o	127	DEL

中文字碼

1.外碼(External code)

- 也稱為輸入碼，是使用者根據某種中文輸入法所輸入的碼或是符號資料。
- 例如:線段法是指電腦中所使用的中文字形，是依每一筆劃的起點、方向以及終點等資料儲存。
- 其外碼種類尚包括有倉頡碼、大易碼、內碼、注音符號碼、電報明碼及三角號碼等。

2.內碼(Internal code)

- 代表真正儲存在電腦內部的碼，包括BCD、EBCDIC、BCDIC、ASCII 等，為英文內碼。
- 而BIG-5 碼、工會碼、王安碼、電報明碼等，為中文內碼。一個內碼通常代表一個英文字或中文字，且長度一定。
- 目前主要的中文碼是以BIG-5 碼為主。其使用兩個位元組表示，其高位元組十六進位值均大於80。

3.交換碼

- 中文內碼有很多種，不同的內碼則無法溝通，當不同的內碼要互傳資料時，要透過大家公認的中間碼來轉換，此中間碼就是交換碼。
- 因此，交換碼為各種內碼間的公共翻譯器，目前所發展的交換碼有兩種：

A. 中文信息交換碼：

- 稱為全漢字交換碼(CCCII；Chinese Character Code for Information Interchange)，為技術學院所研發，用3個位元來表示一中文字，佔用空間較大、效率較低。

B. 通用漢字標準交換碼(CISCII；Chinese Industrial Standard Code for Information Interchange)：

- 為國科會所研發，目前大都使用此一標準，用2個位元來表示一中文字，佔用空間較小、效率較佳。

萬國碼

- UNICODE 即俗稱「萬國碼」的字元編碼標準。由美國萬國碼制訂委員會於1988-1991 年間訂定，涵蓋世界各種不同文字，目前已成為ISO 認證標準(ISO10646)，且發展出兩種編碼方式：UTF-8 與UTF-16。
- UTF-8 為八位元的編碼方式，而UTF-16 即表示十六位元之編碼方式。
- 一般所稱Unicode 係指UTF-16 的形式。因UTF-16 可利用兩個位元組進行編碼，故有多達65,536 種組合，前面128 個符號為ASCII 字元，其餘則為英、中、日、韓文以及其他非英語系國家之38,887 個常用文字。