情報処理の基礎・第 4 回

2013 年 10 月 16 日

今回の内容

4.1 文字列の符号化

日本語や英語で書かれた文章は、文字や数字、句読点などの記号の並びと考えることができます[1]。そこに現れることのできる文字や数字、記号は限られた種類しかありませんから、このような文章も容易にディジタル情報として表すことができます。たとえば、英文に現れる可能性のある文字を考えると、A 〜 Z の大文字 26 種類、a 〜 z の小文字 26 種類、0 〜 9 の数字 10 種類、その他の記号数十種類程度と考えると、基本的には 100 通り程度しかありませんので、7 bit もあれば、1 つの文字を指定することができるはずです。そこで、それぞれの文字に (たとえば) 0 〜 127 までの固有の整数値 (これをその文字の**文字コード**と呼びます) を割り当てておき、この整数値を表すビット列 (たとえば二進法表記) を英文の長さだけ並べることで、英文全体をディジタル情報として表現することができます。

次の表は、ASCII コードと呼ばれる文字コードの割り当て方です。ASCII コードは、計算機や通信機器で文字情報が扱われるときの最も基本的な文字コードの規格として使われています。ASCII コードは、米国で決められた規格ですので、英字や数字、限られた記号などしか表現することができませんが、日本を含め他の地域で使われる文字の文字コードも、多くの場合、この ASCII コードを拡張あるいは若干変更する形で規格が定められています。

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

ASCII コード (右肩の整数が文字コード)

ASCII コードでの文字コードが 0 から 31 までと 127 の文字は**制御文字**と呼ばれ、目に見える文字を表すのではなく、通信手順や表示 (印刷) の書式等を制御するために用いられます。これに対して、英字や数字などの目に見える文字を**図形文字**と呼びます。文字コード 10 の LF は「改行

[1] 文字の大きさや書体の違いは考えないものとします

文字」と呼ばれる制御文字で、行がそこで次の行に移っていることを表します。また、文字コード 32 の SP は「間隔 (スペース)」と呼ばれ、文字と文字の間に置く空白を表すためのものです。SP は制御文字に分類されることも図形文字に分類されることもあります。

符号化の例　例えば、「Ryukoku Univ.」という (スペースや「.」を含めて)13 文字からなる文字列を次のような取り決めに従って符号化してみましょう。

1. 各文字の ASCII コードを 7 bit のビット列 (二進数) で表す。
2. 得られたビット列を左から順に並べて (7 × 文字数)bit のビット列を作る。

まず、文字列「Ryukoku Univ.」の各文字に対応する 7bit のビット列を調べてみると、次の表のようになります。

文字	R	y	u	···	u		U	n	i	v	.
文字コード	82	121	117	···	117	32	85	110	105	118	46
ビット列	1010010	1111001	1110101	···	1110101	0100000	1010101	1101110	1101001	1110110	0101110

よって、この方法で「Ryukoku Univ.」という文字列を符号化すると

$$10100101111001111010101\cdots11101010100000101010111011101101001111011001011 10$$

という 91 bit のビット列になります。

ここでは 1 文字を 7bit で表しましたが、実際には 8bit で表す場合がほとんどです。その場合、同じ文字列は

$$01010010011110010111010101\cdots01110101001000000101010101101110011010010111011000101110$$

という 104 bit (13 B) のビット列に符号化されることになります。

4.2　演習問題

1. 自分の名前のイニシャル (たとえば HN) の各文字を ASCII コード (整数値) で表し、その 2 つの整数値をそれぞれ二進法で 7 bit のビット列に変換して、名、姓の順に並べると 14 bit のビット列になります。こうして得られるビット列を求めなさい。
2. 次のビット列は、ある英単語を構成している文字を ASCII コードで表し、それぞれ二進法で 8 bit のビット列に変換して、左の文字から順に (左から右に) 並べて作ったものである。もとの英単語を求めなさい。

$$01000011011011110110010001100101$$

3. 200 名の人が 100 点満点の試験を受けるときの、それぞれの人の氏名と試験の点数を記録したい。このときのディジタル情報の量は全体で何 byte 必要となるか考えなさい。ただし、氏名はローマ字で書き表し、各文字の ASCII コードを 7bit のビット列にして、30 文字分を

並べて記録する。ローマ字の氏名が 30 文字を越える場合は、31 文字目以降は無視することにし、30 文字に満たない場合は余った部分をスペース (SP) の文字コードで埋めることにする。また、試験の点数は 1 点単位であり、試験に欠席した場合は、そのことが記録されるようにする。1 人分の情報は決まった長さのビット列で記録するものとし、これを 200 名分つなぎ合わせることで全体の情報を記録するものとする。

情報処理の基礎・第 4 回・終り