2.4 文字编码

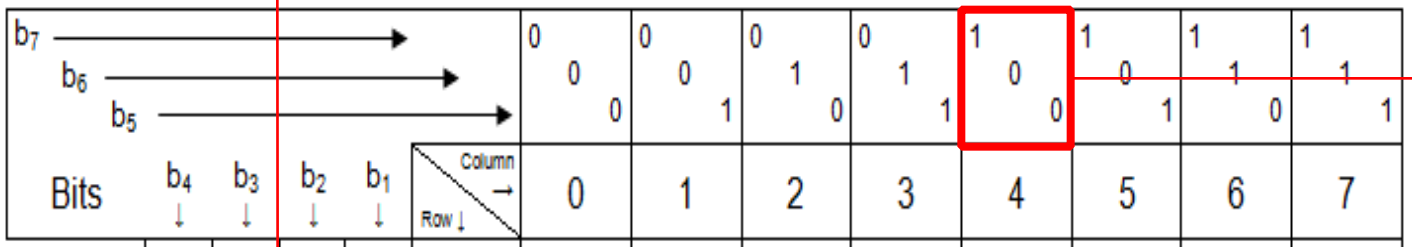
2.4.2 西文字符：

西文字符：指数字、字母以及其它一些符号的总称。西文字符最常用的是ASCII编码。

ASCII码 (American Standard Code for Information Interchange，美国信息交换标准代码)

用7位二进制编码表示128个字符，从0到127

ASCII 编码表



Bits b_4 b_3 b_2 b_1	Column → 0 (000)	1 (001)	2 (010)	3 (011)	4 (100)	5 (101)	6 (110)	7 (111)
Row ↓								
0 0 0 0	0	NUL	DLE	SP	0	@	P	`	p
0 0 0 1	1	SOH	DC1	!	1	A	Q	a	q
0 0 1 0	2	STX	DC2	"	2	B	R	b	r
0 0 1 1	3	ETX	DC3	#	3	C	S	c	s
0 1 0 0	4	EOT	DC4	$	4	D	T	d	t
0 1 0 1	5	ENQ	NAK	%	5	E	U	e	u
0 1 1 0	6	ACK	SYN	&	6	F	V	f	v
0 1 1 1	7	BEL	ETB	'	7	G	W	g	w
1 0 0 0	8	BS	CAN	(	8	H	X	h	x
1 0 0 1	9	HT	EM	)	9	I	Y	i	y
1 0 1 0	10	LF	SUB	*	:	J	Z	j	z
1 0 1 1	11	VT	ESC	+	;	K	[	k	{
1 1 0 0	12	FF	FC	,	<	L	\	l	|
1 1 0 1	13	CR	GS	-	=	M	]	m	}
1 1 1 0	14	SO	RS	.	>	N	^	n	~
1 1 1 1	15	SI	US	/	?	O	_	o	DEL

1000001

字母A的编码

ASCII 包含10个数字

b_7 →				0	0	0	0	1	1	1	1
b_6 →				0	0	1	1	0	0	1	1
b_5 →				0	1	0	1	0	1	0	1
Bits	b_4 ↓	b_3 ↓	b_2 ↓	b_1 ↓	Row ↓ / Column →	0	1	2	3	4	5	6	7
	0	0	0	0	0	NUL	DLE	SP	0	@	P	`	p
	0	0	0	1	1	SOH	DC1	!	1	A	Q	a	q
	0	0	1	0	2	STX	DC2	"	2	B	R	b	r
	0	0	1	1	3	ETX	DC3	#	3	C	S	c	s
	0	1	0	0	4	EOT	DC4	$	4	D	T	d	t
	0	1	0	1	5	ENQ	NAK	%	5	E	U	e	u
	0	1	1	0	6	ACK	SYN	&	6	F	V	f	v
	0	1	1	1	7	BEL	ETB	'	7	G	W	g	w
	1	0	0	0	8	BS	CAN	(	8	H	X	h	x
	1	0	0	1	9	HT	EM	)	9	I	Y	i	y
	1	0	1	0	10	LF	SUB	*	:	J	Z	j	z
	1	0	1	1	11	VT	ESC	+	;	K	[	k	{
	1	1	0	0	12	FF	FC	,	<	L	\	l	|
	1	1	0	1	13	CR	GS	-	=	M	]	m	}
	1	1	1	0	14	SO	RS	.	>	N	^	n	~
	1	1	1	1	15	SI	US	/	?	O	_	o	DEL

ASCII 包含34个控制字符

Bits b_4 b_3 b_2 b_1	Row ↓ / Column →	b_7 b_6 b_5 = 0 0 0	0 0 1	0 1 0	0 1 1	1 0 0	1 0 1	1 1 0	1 1 1
		0	1	2	3	4	5	6	7
0 0 0 0	0	NUL	DLE	SP	0	@	P	`	p
0 0 0 1	1	SOH	DC1	!	1	A	Q	a	q
0 0 1 0	2	STX	DC2	"	2	B	R	b	r
0 0 1 1	3	ETX	DC3	#	3	C	S	c	s
0 1 0 0	4	EOT	DC4	$	4	D	T	d	t
0 1 0 1	5	ENQ	NAK	%	5	E	U	e	u
0 1 1 0	6	ACK	SYN	&	6	F	V	f	v
0 1 1 1	7	BEL	ETB	'	7	G	W	g	w
1 0 0 0	8	BS	CAN	(	8	H	X	h	x
1 0 0 1	9	HT	EM	)	9	I	Y	i	y
1 0 1 0	10	LF	SUB	*	:	J	Z	j	z
1 0 1 1	11	VT	ESC	+	;	K	[	k	{
1 1 0 0	12	FF	FC	,	<	L	\	l	|
1 1 0 1	13	CR	GS	-	=	M	]	m	}
1 1 1 0	14	SO	RS	.	>	N	^	n	~
1 1 1 1	15	SI	US	/	?	O	_	o	DEL

ASCII 包含52个英文字母

b_7 →					0	0	0	0	1	1	1	1
b_6 →					0	0	1	1	0	0	1	1
b_5 →					0	1	0	1	0	1	0	1
Bits	b_4 ↓	b_3 ↓	b_2 ↓	b_1 ↓	Row ↓ / Column →	0	1	2	3	4	5	6	7
	0	0	0	0	0	NUL	DLE	SP	0	@	P	`	p
	0	0	0	1	1	SOH	DC1	!	1	A	Q	a	q
	0	0	1	0	2	STX	DC2	"	2	B	R	b	r
	0	0	1	1	3	ETX	DC3	#	3	C	S	c	s
	0	1	0	0	4	EOT	DC4	$	4	D	T	d	t
	0	1	0	1	5	ENQ	NAK	%	5	E	U	e	u
	0	1	1	0	6	ACK	SYN	&	6	F	V	f	v
	0	1	1	1	7	BEL	ETB	'	7	G	W	g	w
	1	0	0	0	8	BS	CAN	(	8	H	X	h	x
	1	0	0	1	9	HT	EM	)	9	I	Y	i	y
	1	0	1	0	10	LF	SUB	*	:	J	Z	j	z
	1	0	1	1	11	VT	ESC	+	;	K	[	k	{
	1	1	0	0	12	FF	FC	,	<	L	\	l	|
	1	1	0	1	13	CR	GS	-	=	M	]	m	}
	1	1	1	0	14	SO	RS	.	>	N	^	n	~
	1	1	1	1	15	SI	US	/	?	O	_	o	DEL

ASCII 包含32个标点符号与运算符号

b_7 →					0	0	0	0	1	1	1	1
b_6 →					0	0	1	1	0	0	1	1
b_5 →					0	1	0	1	0	1	0	1
Bits	b_4 ↓	b_3 ↓	b_2 ↓	b_1 ↓	Row ↓ / Column →	0	1	2	3	4	5	6	7
	0	0	0	0	0	NUL	DLE	SP	0	@	P	`	p
	0	0	0	1	1	SOH	DC1	!	1	A	Q	a	q
	0	0	1	0	2	STX	DC2	"	2	B	R	b	r
	0	0	1	1	3	ETX	DC3	#	3	C	S	c	s
	0	1	0	0	4	EOT	DC4	$	4	D	T	d	t
	0	1	0	1	5	ENQ	NAK	%	5	E	U	e	u
	0	1	1	0	6	ACK	SYN	&	6	F	V	f	v
	0	1	1	1	7	BEL	ETB	'	7	G	W	g	w
	1	0	0	0	8	BS	CAN	(	8	H	X	h	x
	1	0	0	1	9	HT	EM	)	9	I	Y	i	y
	1	0	1	0	10	LF	SUB	*	:	J	Z	j	z
	1	0	1	1	11	VT	ESC	+	;	K	[	k	{
	1	1	0	0	12	FF	FC	,	<	L	\	l	|
	1	1	0	1	13	CR	GS	-	=	M	]	m	}
	1	1	1	0	14	SO	RS	.	>	N	^	n	~
	1	1	1	1	15	SI	US	/	?	O	_	o	DEL

•按照上面提供的ASCII码，就可以把字符串“code”表示为：

c	o	d	e
01100011	01101111	01100100	01100101

•利用ASCII标准对字符串“1+2”进行编码，可以表示为：

1	+	2
00110001	00101011	00110010

控制字符：34个（0～32，127）；

图形字符（普通字符）：94个。

‘0’～‘9’　30H～39H　48～57

‘A’～‘Z’　41H～5AH　65～90

‘a’～‘z’　61H～7AH　97～122

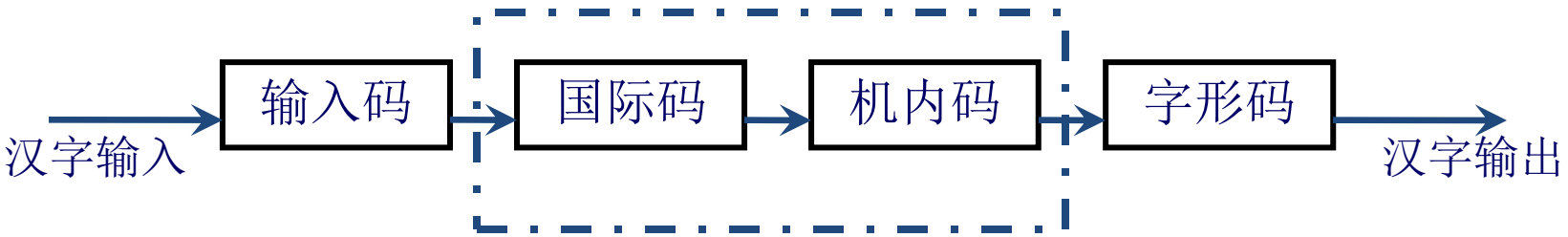
◆ 小写字母的编码比对应大写字母的编码大32；

例如：

“a”字符的编码为1100001，对应的十进制数是97；

“A”字符的编码为1000001，对应的十进制数是65；

2.4.3 汉字编码



(1) 汉字外码

外码也叫输入码，主要解决如何将每个汉字变成可以直接从键盘输入的代码。目前常用的输入法主要是音码和形码两类。

音码类： 全拼、双拼、微软拼音、自然码和智能ABC等

形码类： 五笔字型法、郑码输入法等 。

2.4.3 汉字编码

(2) 汉字国标码(GB2312－80)

汉字国标码是1980年发布的《中华人民共和国标准信息交换编码》，代号为GB2312-80，简称国标码。

国标码是二字节码，既用二个字节的低7位进行二进制数编码来表示一个汉字,每个字节的最高位置都是0。

• 区位码

汉字94×94的矩阵，即94个区和94个位，由区号和位号构成汉字的区位码 。

中：5448

华：2710

区号	位号

• 汉字的国标码与区位码的关系：

每个汉字的区号和位号各加32(20H)就构成了国标码

加32的原因：为了与ASCII码兼容，每个字节值大于32（0～32为非图形字符码值）

每个汉字的编码占两个字节, 使用每个字节的低7位，共14位

(3)汉字机内码

汉字在设备或信息处理系统内部最基本的表达形式。为了在计算机内部能够区分是汉字编码还是ASCII码，将国标码每个字节最高位设置为1(80H).

	区位码	国标码	机内码
中	(36 30)H	(56 50)H=(01010110 01010000)B	(11010110 11010000)B=(D6 D0)H
华	(1B0A)H	(3B 2A)H=(00111011 00101010)B	(10111011 10101010)B=(BB AA)H

三种码之间关系：

汉字机内码=汉字国标码+80 80H=区位码+A0 A0H

国标码=区位码+20 20H

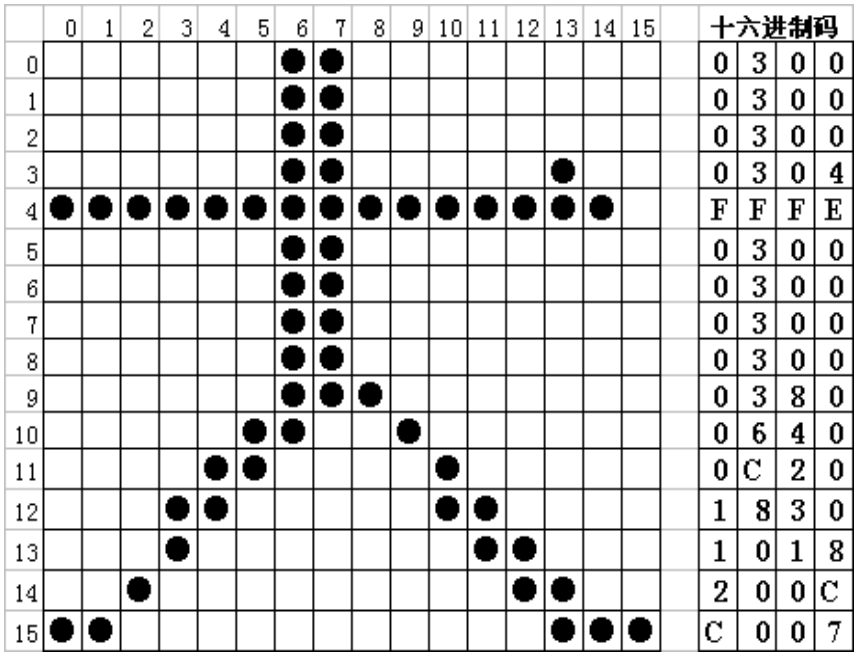
(4) 汉字字形码

点阵：汉字字形点阵的代码，有16×16、24×24、32×32 、48×48等编码、存储方式简单、无需转换直接输出放大后产生的效果差。

矢量：存储的是描述汉字字形的轮廓特征

矢量方式特点正好与点阵相反

“大”字的16×16点阵及代码



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	十六进制码			
0							●	●									0	3	0	0
1							●	●									0	3	0	0
2							●	●									0	3	0	0
3							●	●						●			0	3	0	4
4	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●		F	F	F	E
5							●	●									0	3	0	0
6							●	●									0	3	0	0
7							●	●									0	3	0	0
8							●	●									0	3	0	0
9							●	●	●								0	3	8	0
10						●	●			●							0	6	4	0
11					●	●					●						0	C	2	0
12				●	●						●	●					1	8	3	0
13				●								●	●				1	0	1	8
14			●										●	●			2	0	0	C
15	●	●												●	●	●	C	0	0	7

(5) 几种常见的汉字编码

•Unicode字符集

另一国际标准：采用双字节编码统一地表示世界上的主要文字。目前的Unicode字符分为17组编排。UTF-8、UTF-16、UTF-32是常用的几组编码方案。

•UTF-8编码

UTF-8的特点是对不同范围的字符使用不同长度的编码，0～127之间的码字都使用一个字节存储，超过128的码字使用2～4个字节存储。

•UTF-16编码

UTF-16中的字符，要么用2个字节表示，要么用4个字节表示。

(5) 几种常见的汉字编码

- GBK码

GBK等同于UCS的新的中文编码扩展国家标准，2字节表示一个汉字。

- BIG5编码

台湾、香港地区普遍使用的一种繁体汉字的编码标准，包括440个符号，一级汉字5 401个、二级汉字7 652个，共计13 060个汉字。