计算机导论

山东建筑大学
计算机学院
秦松

Chapter 2

数据的表示

数据

数据（data）在计算机科学中是指所有能输入到计算机并被计算机程序处理的符号的介质的总称。

数据处理

- 围绕着数据所做的工作均称为数据处理。
- 数据处理是指对数据的收集、组织、整理、加工、存储和传播等工作。

数据处理

数据处理分为3类：

1) 数据管理：收集信息、将信息用数据表示并按类别组织保存；
   - 组织和保存数据；
   - 进行数据维护；
   - 提供数据查询和数据统计功能。

2) 数据加工：对数据进行变换、抽取和运算；

3) 数据传播：在空间或时间上以各种形式传播信息，而不改变数据的结构、性质和内容；

数据的特征

数据受数据类型和取值范围的约束。

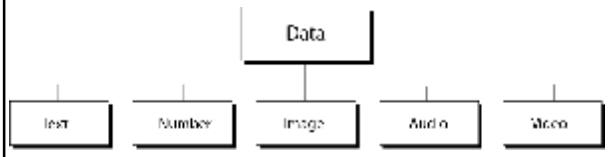
取值范围：数据的值域。

如：学生“性别”的值域是 { “男”，“女” }；

数据类型

不同类型的数据

数据以多种表现形式出现



工程程序: number. 例1

文字处理程序: text. 例2

图像处理程序: images. 例3

多媒体

注:

计算机业使用术语"多媒体"来定义包含数字、文本、图像、音视频的信息。

例：片断 1'04"

计算机内部的数据

存储数据

最好的数据存储方式是电子信号，以其出现和消失的特定方式存储。这意味着可以以两种状态之一存储数据（0 or 1）。

尽管数据只能以一种形式（二进制）存储在计算机内部，但在计算机外部则表现为多种形式。

数据组织

计算机外部所有的数据类型要转换成统一的表示法然后被计算机储存并且可以被还原。

这种通用的格式叫做位模式(bit pattern).

为了方便存储数据被组织成许多小的单元，再由这些小的单元组成更大的单元。

位（bit，binary digit）是存储在计算机中的最小数据单元：它是0或1。位代表设备的某一种状态。

位模式

单个位不能解决数据表示问题。

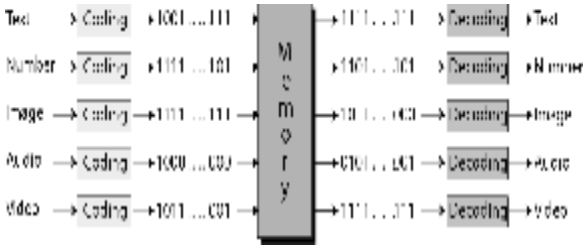
为了表示数据的不同类型，应该使用位模式，一个序列，有时也被称之为位流。

Bit pattern

1 0 0 0 1 0 1 0 1 0 1 1 1 1 1

位模式

数据输入机器时被编码，输出时被解码



字节

长度为8的位模式叫字节byte.

$1K=2^{10}=1024$ byte，

$1M=1024K=2^{10}2^{10}$ byte

$1G=1024M=2^{10}2^{10}2^{10}$ byte

表示数据

Question

问题

在一种语言中位模式需要多少位表示一个符号?

他取决于该语言集中有多少不同的符号。位模式的长度和符号数量是对数关系：

如果需要2个符号，位模式的长度是1位($\log_2 2=1$)。

2位的位模式能表示成四种形式:

00 01 10 11

如果需要4个符号，长度是2位($\log_2 4=2$) .

3位的位模式:

000 001 010 011 100 101 111

代码

不同的位模式集合被设计用于表示文本符号。每一个集合称为代码，表示符号的过程叫编码。

常用代码：

ASCII码 美国国家标准协会（ANSI）美国信息交换标准码

ASCII 码的一些突出特点：

- 使用 7 位模式，范围从0000000 到 1111111。
- 有31种控制（不可打印）字符。
- 数字字符(0 to 9)在字母字符之前。
- 大写字符(A…Z)编码在小写字符(a…z)之前。

ASCII码

表 ASCII 码字符集

高3位 / 低4位	000	001	010	011	100	101	110	111
0000	NUL	DC0	SP	0	@	P	-	p
0001	SOH	DC1	!	1	A	Q	a	q
0010	STX	DC2	"	2	B	R	b	r
0011	ETX	DC3	#	3	C	S	c	s
0100	EOT	DC4	$	4	D	T	d	t
0101	ENQ	NAK	%	5	E	U	e	u
0110	ACK	SYN	&	6	F	V	f	v
0111	BEL	ETB	,	7	G	W	g	w
1000	BS	CAN	)	8	H	X	h	x
1001	HT	EM	(	9	I	Y	I	y
1010	LF	SUB	*	:	J	Z	J	z
1011	VT	ESC	+	;	K	[	k	{
1100	FF	FS	,	<	L	\	l	|
1101	CR	GS	-	=	M	]	m	}
1110	SO	RS	.	>	N	↑	n	~
1111	SI	US	/	?	O	←	o	DEL

常用代码

扩展 ASCII是每一个位模式统一为 1 字节（8位），ASCII位模式通过在左边增加额外的 0 来扩充。(00000000 to 01111111)

一些制造商希望再附加128个字符，但因为能制定出统一标准的代码集，未成行

EBCDIC

早期IBM开发的扩充的二进制编码十进制交换码，使用8位模式，最多表示256个字符，只用在IBM大型机上

常用代码

Unicode

由(一开始大多是美国的)硬件和多语言软件制造商组成的协会组织的 Unicode 项目.定义一个16位65536个符号。它覆盖了美国、欧洲、中东、非洲、印度、亚洲和太平洋的语言，以及古文和专业符号。Unicode 允许交换、处理和显示多语言文本以及公用的专业和数学符号。它希望能够解决多语言的计算，如不同国家的字符标准，但并不是所有的现代或古文都能够获得支持。

JAVA使用，Windows使用了前256个字符的一个变化版本

ISO国际标准化组织

使用32位模式的代码，能表示4294967296个符号，可表示任何符号

汉字字符在机内如何表示

汉字是一种特殊的字符，同样采用编码的形式在计算机内表示和存储它。
《信息交换用汉字编码字符集 — 基本集》即国家标准 GB2312 - 80 就是这样的编码表。
汉字编码表比ASCII编码表要大得多，它由94×94的表构成，即有94行，94列。每一行称为一个“区”，共有94区，编号为第01区、第02区、…、第94区；每一列称为一个“位”，共94位，编号为第01位、第02位、…、第94位。
在港澳台和境外地区华人界使用BIG-5编码标准。
汉字的表示就使用该编码表提供的编码，即用“区码”和“位码”作为汉字的编码，“区码”为高位（在左），“位码”为低位（在右）。

区位码表（局部）

"南京"二字，根据汉字编码表可知其编码分别为：
[南] 十进制码是：36 47　二进制码是：0100100 0101111
[京] 十进制码是：30 09　二进制码是：0011110 0001001

	01	02	03		09		47		94
01									
02									
03									
30					**京**				
36							**南**		
94									

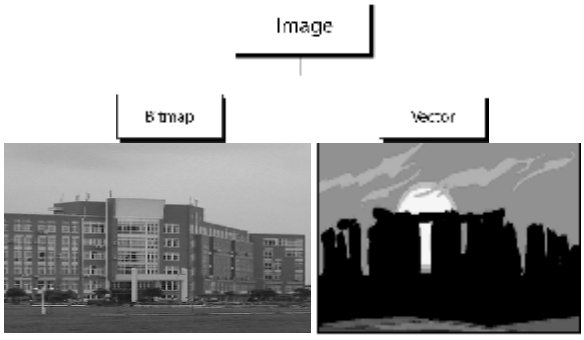
其他表示

国标码
在区位码的基础上产生，方法是分别在“区”码和“位码”上各加“32”（即二进制00100000）得到。
南 1000100 1001111 京 0111110 0101001

机内码
在国标码的基础上产生，分别在"高位"码和"低位"码上个各加"128"（即二进制10000000）得到。
南 11000100 11001111 京 10111110 10101001

图像

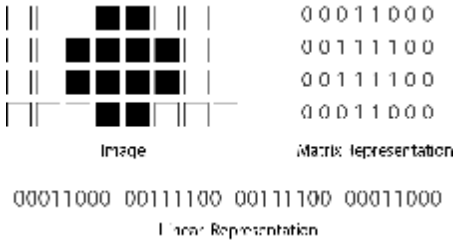
图像表示方法：位图图形和矢量图形



位图图形

图像被分成像素矩阵, 每一个像素是一个小点。 像素的大小取决于分辨率

黑-白图像的位图图形表示方法

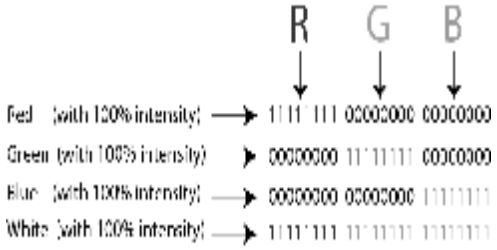


位图图形

图像灰阶：图像如果不是纯黑纯白得像素组成，可以增加位模式的长度表示灰色。
在256-灰阶图像，每个像素可能是白、黑或另２５４灰阶中的一个。

位图图形

表示彩色像素：每一种彩色像素被分解为红绿蓝，用一个8位模式表示其强度。



分辨率

256-灰阶图像

640×480=307200 pixel

640×480×8=307200 ×8 bits

640×480×8/8=307200 bytes

分辨率

表示彩色像素

16色、256色、真彩色图形的存储空间:

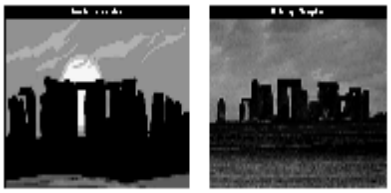
16色:640×480×4/8

256色:640×480×8/8

真彩色:640×480×24/8

矢量图像

矢量图像表示方法不存储位模式。
图像分解为曲线和直线的组合，其中每一曲线或直线有数学公式表示。
这些公式的组合存储在计算机中。
当打印显示图像时，系统根据新尺寸和公式计算图像。



图像文件

位图图像文件
.bmp、.pcx、.tif、.jpg和.gif

使用位图软件修改位图文件
Microsoft Paint、Adobe Photoshop等

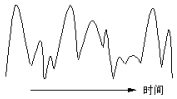
矢量图像文件
.wmf、.dxf、.mgx、.cgm

矢量图像软件
CorelDraw、Micrographx Designer等

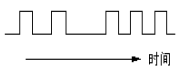
两类信号形式

信号（signal） 一般可分为两类：

模拟信号（analog signal）连续变化值（如数学中之实数）。自然界产生的物理量一般均为模拟信号。

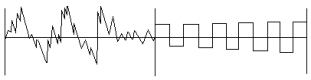


数字信号（digital signal）离散变化值（如数字中之整数）。计算机内部传输和处理的均为矩形脉冲形式的数字信号（1和0）。



音频

音频表示声音,没有标准,音频模拟数据转换成数字数据



转换步骤:

对模拟信号采样. 采样就是在相等的间隔来测量信号值。

量化采样值。量化就是给采样值分配值（从规定的值集中）.

量化值转换为位模式。

存储位模式.

视频

视频是图像在时间上的表示（帧）.

电影就是一系列帧,每一帧就是一幅静态图像,一张张连续播放.像小时候的小人书电影

存储视频即图像,通常被压缩RM,WMV,MPEG….