

扩展第5讲 机器是怎样表示文字声音与图像的 -编码与解码

战 德 臣

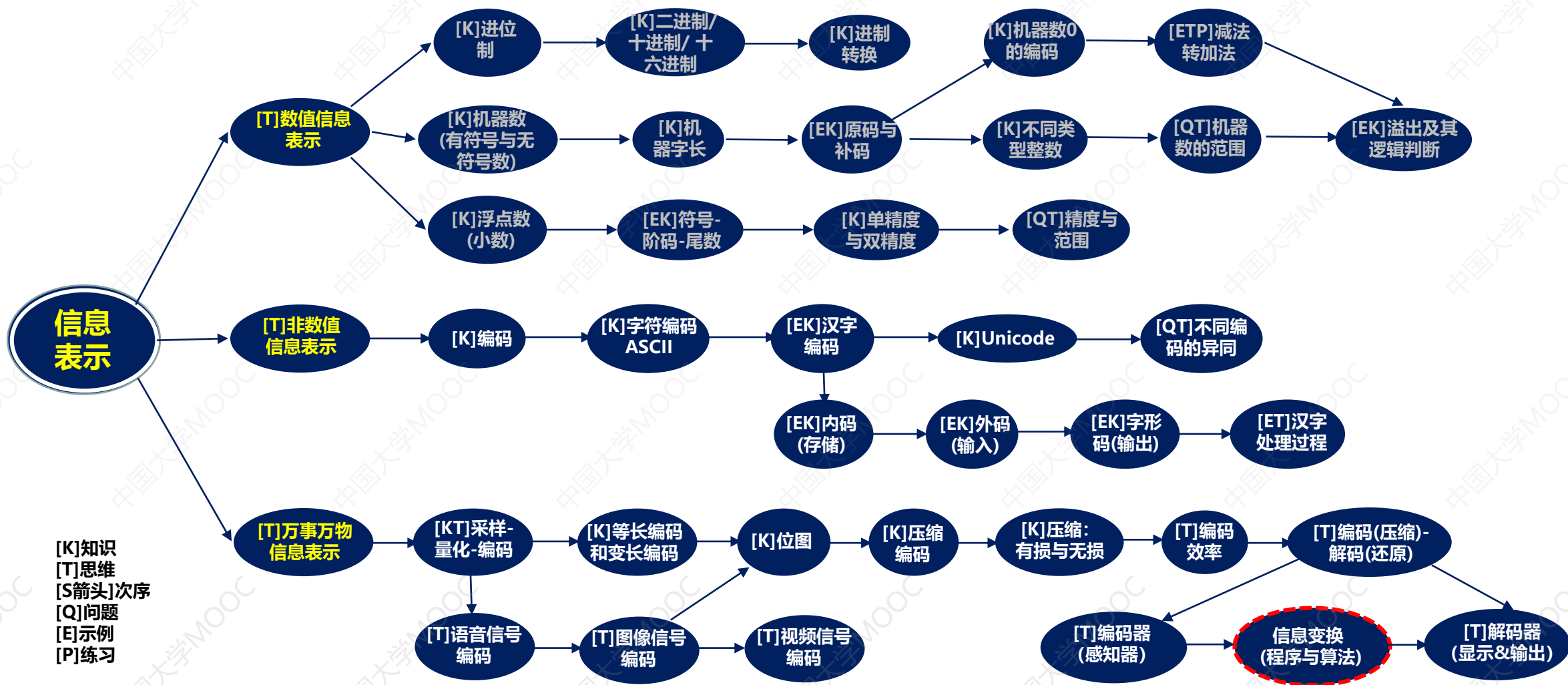
哈尔滨工业大学计算学部教学委员会主任
国家教学名师

18686783018, dechen@hit.edu.cn

信息在计算机中的基本表示方法

2

概览



为什么需要编码

3

非数值性信息的表达手段--编码

非数值性信息可以用编码表示

◆ **编码**：编码是以若干位数码或符号的不同组合来表示非数值性信息的方法，它是人为地将若干位数码或符号的每一种组合指定一种唯一的含义。

例如：0----男，1----女

再如：001--星期一；010--星期二；011--星期三；100--星期四；101--星期五；110--星期六；000--星期日

编码的主要特征：

- ◆ **唯一性**：每一种组合都有确定的唯一性的含义
- ◆ **公共性**：所有相关者都认同、遵守、使用这种编码
- ◆ (人使用编码)：**易于记忆/便于识认**
- ◆ (机器使用编码)：(1)有一定规律，便于机器解读；(2)利用0-1编码；(3)效率：存储效率(以最少的位数来编码最多的信息，最少的位数即占用最少的存储空间)和编码/解码算法的效率(编码存储与解码还原的效率)。
- ◆ **编码是典型的计算思维**，利用0-1编码求解问题。

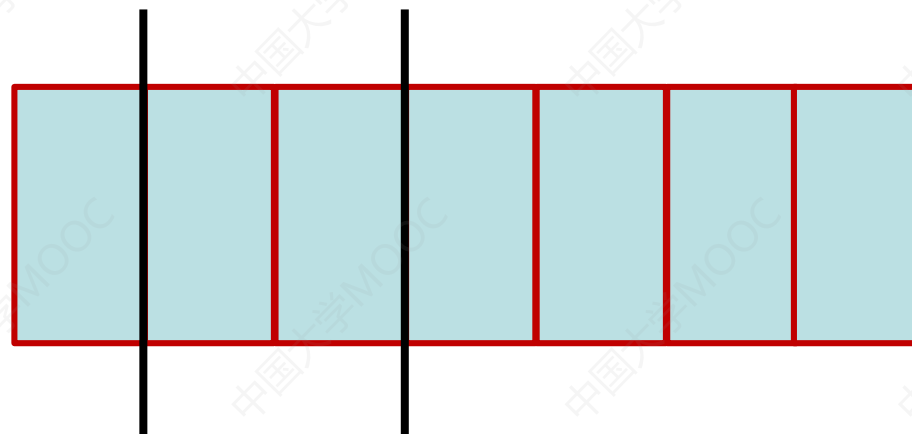
为什么需要编码

编码是典型的计算思维

【示例】有一位雇主，雇佣一位员工为自己完成一项7天能够完成的任務。雇主答应总计支付一根金条作为报酬，员工要求每天支付1/7的工资，雇主答应了。问：雇主怎样在金条上切两刀，保证每天正好能支付员工1/7的工资。

【思维】

- (1) 切两刀，形成3块，即3个数字（3个数字都必须是1/7的倍数）
- (2) 3个数字通过组合能否组合出1-7的每一个。
- (3) 按2的幂次方来切：三块分别为 $2^0, 2^1, 2^2$ 。
- (4) 由2的幂次方可组合出1-7的每一个。



7	6	5	4	3	2	1
D7	D6	D5	D4	D3	D2	D1
111	110	101	100	011	010	001

二进制编码求解问题是一种基本思维

关于编码

5

先看几种常规的等长编码

- ASCII码
- 汉字编码
- Unicode编码

ASCII码

英文字母符号编码--ASCII码

ASCII码是英文字母与符号的0-1型编码方法，是用7位0和1的不同组合来表示10个数字、26个英文大写字母、26个英文小写字母及其一些特殊符号的编码方法，是信息交换的标准编码。为处理方便，通常采用8位0和1编码，其中最高位为0。

ASCII码为什么
要用8位来
编码呢？

◆ASCII码：American Standard Code for Information Interchange

$B_7 B_6 B_5 B_4 B_3 B_2 B_1 B_0$

0 x x x x x x x
0 0 1 1 0 0 0 1 “1”
0 1 0 0 1 1 1 0 “N”

每8位为一个字符，最高位为0
✓41H ~ 5AH: “A” ~ “Z”
✓61H ~ 7AH: “a” ~ “z”
✓30H ~ 39H: “0” ~ “9”
✓0AH: 换行符号LF
✓0DH: 回车符号CR

$b_7 b_6 b_5 b_4$ $b_3 b_2 b_1 b_0$	$B_7 B_6 B_5 B_4$ 000	001	010	011	100	101	110	111
0000	NUL	DLE	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

由ASCII码看键盘与显示器

编码与解码示例：键盘-ASCII-显示/输出

信息	存储	解析规则
We are students	01010111 01100101 00100000 01100001 01110010 01100101 00100000 01110011 01110100 01110101 01100100 01100101 01101110 01110100 01110011	0/1串按8位分隔一个字符，查找ASCII码表映射成相应符号

b ₇ b ₆ b ₅ b ₄ b ₃ b ₂ b ₁ b ₀	000	001	010	011	100	101	110	111
0000	NUL	DLE	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

八位0/1码绑定语义：
一个字母或数字

信息表示规范/标准：
ASCII码

解码器：读取文件，8位
分隔一个字符，将8位
ASCII码转换成字符字
型码送显示器显示



信息采集：键盘(通过按键
位置识别出所按的符号)

编码器：将符号转换
成ASCII码存储/8
位一个字符

信息：ASCII码
0/1存储的文件
(.txt)

键盘的实
现机理是
怎样的？

应用程序接收的
是操作系统传递
过来的ASCII码

【1】字符发生器；
【2】光栅/点阵-显
示用内存

显示的实
现机理是
怎样的？

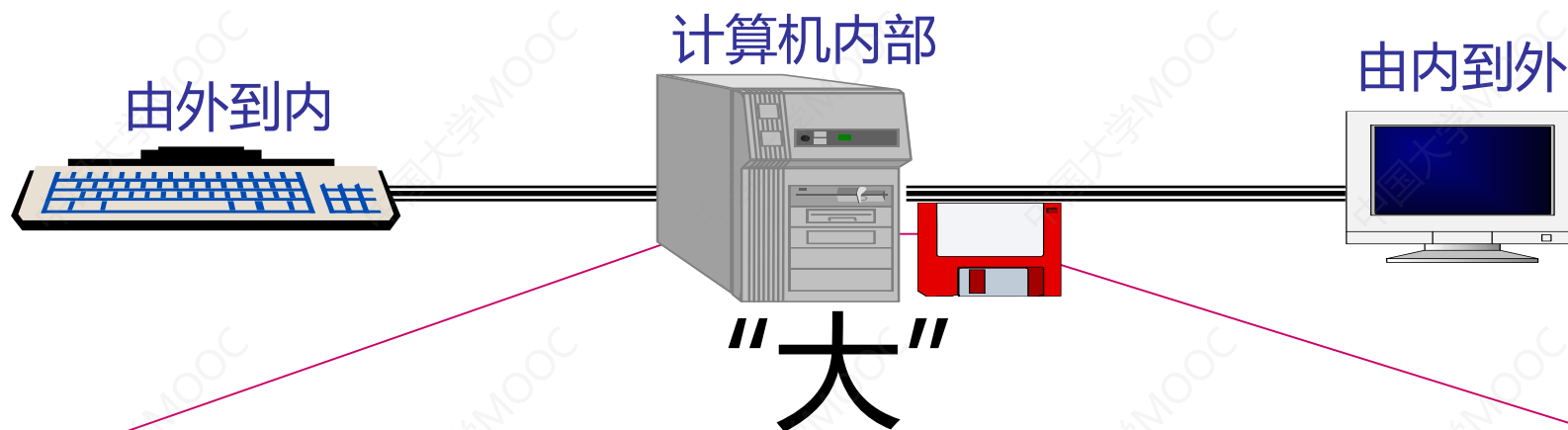
信息显示/
打印

汉字内码

8

汉字在机器内部的存储—汉字内码

汉字在计算机内部采用汉字内码存储，**汉字内码**是一两字节且最高位均为1的0-1型编码。



汉字编码标准

- GB2312-80
- GBK-95
- GB18030-2000
- GB18030-2005

用0和1编码汉字,每个汉字在计算机内部由 2个字节表示

	b_7	b_6	b_5	b_4	b_3	b_2	b_1	b_0	b_7	b_6	b_5	b_4	b_3	b_2	b_1	b_0
国标码	0	0	1	1	0	1	0	0	0	1	1	1	0	1	1	1
(机)内码	1	0	1	1	0	1	0	0	1	1	1	0	1	1	1	1

单字节操作
系统—西
文操作系统

两字节操作
系统—中
文操作系统

两字节能编
码多少个汉
字，够用吗？

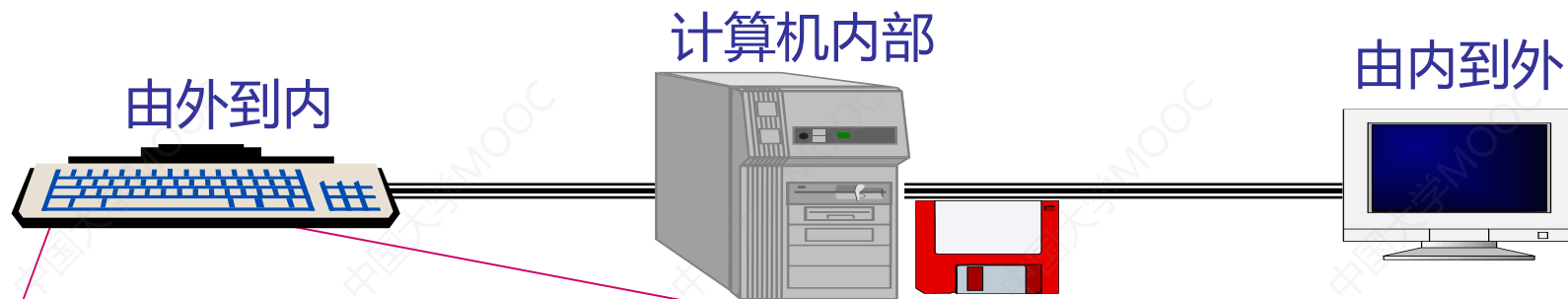
怎样区分一个
字节是ASCII码
还是汉字编码？

汉字外码/输入码

9

汉字如何输入到计算机中——汉字输入码/外码

汉字外码是用键盘上的字母符号编码每一汉字的编码, 它使人们通过键入字母符号代替键入汉字



输入码有若干: 拼音码、字型码、区位码... ..

拼音码: xing

双拼码: x;

其中, 'x' 表声母x,而 ';' 表韵母ing

五笔字型码: gajf

其中,g表字根 "一",a表开下的草字头,j表右侧立刀,f表下面土字

“型”

为什么会有那么多汉字编码?

汉字外码编码要解决什么问题?

一个编码能否唯一对应一个汉字?

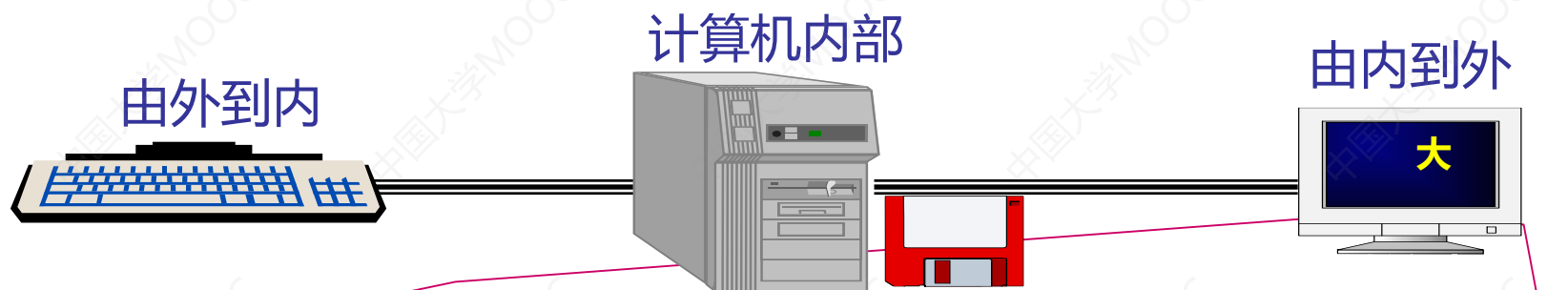
怎样方便人们记忆?

怎样减少汉字输入的按键次数?

汉字字形码

汉字如何显示和打印—汉字字形码

汉字字形码是用0和1编码无亮点和有亮点像素,形成汉字字形的一种编码,依据字形码通过显示器或打印机输出汉字。



用0和1编码无亮点和有亮点形成字形信息, 便于显示... ..

汉字字形码是一种字模点阵码。也有不同的处理汉字点阵信息的编码，如向量编码等

“大”

怎样编码汉字的字形？

字模点阵码

汉字向量码 / 矢量码

Unicode编码

11

世界所有字符/文字的统一编码：Unicode

Unicode是可以容纳世界上所有文字和符号的字符编码方案，用数字0到0x10FFFF的范围来映射所有的字符。具体处理时，将前述唯一确定的码位按照不同的编码方案映射为相应的编码，有**UTF-8**、**UTF-16**、**UTF-32**等几种编码方案。

Unicode: $b_{15}b_{14}b_{13}b_{12}b_{11}b_{10}b_9b_8b_7b_6b_5b_4b_3b_2b_1b_0$

【UTF-8】是以字节（8位比特）为单位对Unicode字符进行编码，对不同范围的字符使用不同长度的编码：

(1) 0x000000至0x00007F区间的符号（此区间的字符为标准的ASCII码字符），用一字节编码，即0x00至0x7F，将高位的0均省略，与标准的ASCII码完全相同。(2) 0x000080至0x0007FF区间的符号，用两字节编码，即**110**xxxxx**10**xxxxxx。(3) 0x000800至0xFFFF区间的符号(中文汉字通常位于此区间),用三字节编码**110** $b_{15}b_{14}b_{13}b_{12}$ **10** $b_{11}b_{10}b_9b_8b_7b_6$ **10** $b_5b_4b_3b_2b_1b_0$ 。(4) 其它区间，则分别用四字节、五字节和六字节编码，最大长度是6个字节。

【UTF-16】是以字（16位比特）为单位对Unicode字符进行编码，在0x000000至0x00FFFF区间的符号（国标汉字和ASCII码字符通常位于此区间）其后16位就是其UTF-16编码。**【UTF-32】**是以双字（32位比特）为单位对Unicode字符进行编码。

“大”的Unicode码为01011001 00100111。则：

“大”的UTF-8编码为**110** 0101 **10**100100 **10**100111，即0x E5 A4 A7。

“大”的UTF-16编码为 01011001 00100111，即0x 59 27。

“灯”的Unicode码为01110000 01101111。则：

“灯”的UTF-8编码为**110** 0111 **1000** 0001 **10**10 1111，即0x E7 81 AF。

“灯”的UTF-16编码为 01110000 01101111，即0x 70 6F。

GB18030和
Unicode能
对应吗？怎
样对应？

关于编码

12

再看一种不等长编码（霍夫曼编码）——如何提高编码效率

- **霍夫曼编码**是依据对象出现的概率分布来确定对象编码长度的一种变长编码方式，在保证编码不存在二义性的前提下，实现总体编码长度的最小化。
- 对出现概率大的对象赋予较短的编码，而对出现概率小的对象赋予较长的编码。

采采芣苢 薄言采之 采采芣苢 薄言有之
采采芣苢 薄言掇之 采采芣苢 薄言捋之
采采芣苢 薄言袺之 采采芣苢 薄言襁之

- 等长编码：全诗共 48 字，不重复汉字有 11 个，需采用 4 位二进制编码，总长度为 $4 \times 48 = 192$ 比特。
- 不等长编码(霍夫曼编码)：全诗 48 字可用 143 个比特完成编码，平均码长为 2.979。

采	0000	采	00
芣	0001	芣	010
苢	0010	苢	011
薄	0011	薄	100
言	0100	言	101
之	0101	之	110
掇	0110	掇	11100
袺	0111	袺	11110
有	1000	有	11111
捋	1001	捋	111010
襁	1010	襁	111011

这个编码是怎样产生的呢？

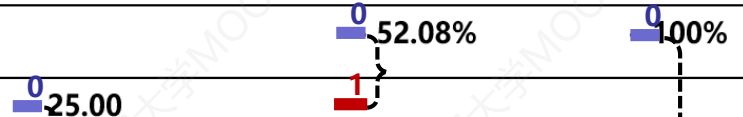
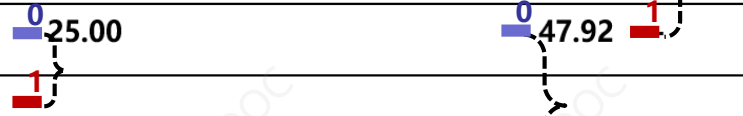
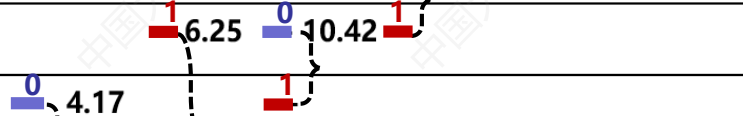
关于编码

13

霍夫曼编码【示例】

采采芣苢 薄言采之 采采芣苢 薄言有之
采采芣苢 薄言掇之 采采芣苢 薄言捋之
采采芣苢 薄言桔之 采采芣苢 薄言櫨之

霍夫曼编码：(1)计算每个汉字出现的概率。(2)按照如下规则进行合并相加：即不断将全表中尚未使用的最小的两个概率数字求和归并，直到得到100%。(3)反向依据归并路径编码。从右往左，蓝色路径编码 0，红色路径编码 1，走完路径即可得到每一个汉字的编码。

汉字	字频	概率		编码	长度
采	13	27.08%		00	2
芣	6	12.50%		010	3
苢	6	12.50%		011	3
薄	6	12.50%		100	3
言	6	12.50%		101	3
之	6	12.50%		110	3
掇	1	2.08%		11101	5
桔	1	2.08%		11110	5
有	1	2.08%		11111	5
捋	1	2.08%		111000	6
櫨	1	2.08%		111001	6

关于编码

14

是否还能提高编码效率—压缩编码

采采芣苢 薄言采之 采采芣苢 薄言有之
采采芣苢 薄言掇之 采采芣苢 薄言捋之
采采芣苢 薄言桔之 采采芣苢 薄言禴之

等长编码需4位编码总计需192比特
霍夫曼不等长编码总计需143比特
是否还能更少?

重复短语, 若建立字典:

A: 采采芣苢 薄言
B: 采
C: 之 采采芣苢 薄言
D: 有
E: 掇
F: 捋
G: 桔
H: 禴
I: 之

则全诗可以从 48 个汉字字符
压缩为 13 个英文字符:

A B C D C E C F C G C H I



字母	字频	概率		编码	长度
C	5	38.46%		1	1
A	1	7.69%	0 15.38 0 30.77 0 61.53 0	0000	4
B	1	7.69%	1	0001	4
D	1	7.69%	0 15.38 1	0010	4
E	1	7.69%	1	0011	4
F	1	7.69%	0 15.38 0 30.77 1	0100	4
G	1	7.69%	1	0101	4
H	1	7.69%	0 15.38 1	0110	4
I	1	7.69%	1	0111	4

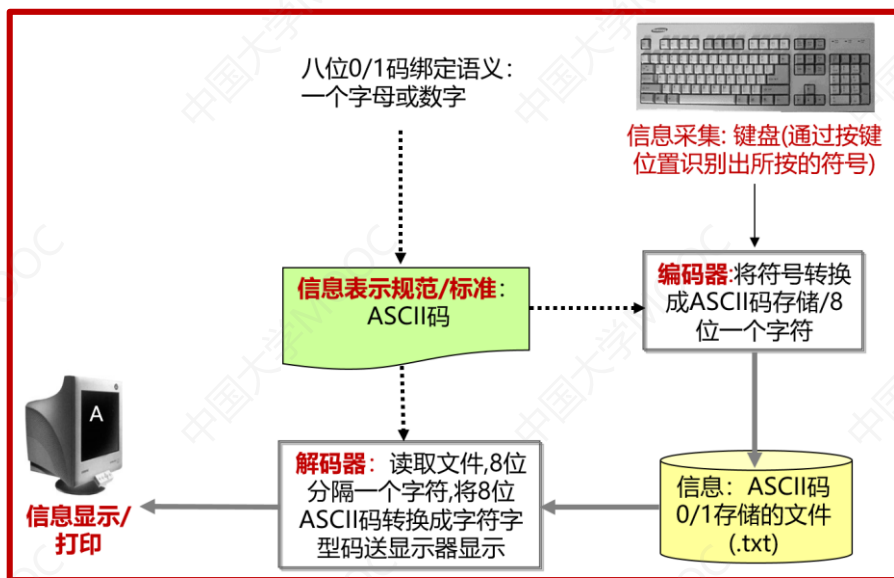


再用霍夫曼编码存储, 则仅需要 37 个比特

0000 0001 1 0010 1 0011 1 0100 1 0101 1 0110 0111

万事万物之符号化与编码

协议与编码器和解码器——一种抽象与自动化机制



- (研究对象的)信息绑定语义的方法: 0/1码绑定或字母符号绑定--何种语义。
- 一般而言, “协议”是为交流信息的双方(计算机)能够正确实现信息交流而建立的一套规则、标准或约定。

(研究对象的)协议/语言/标准

感知器/传感器

(研究对象的)信息采集

编码器

录音机、照相机、摄像机、是否就是一种编解码器呢?

按协议/语言/标准表示和存储的信息

编译器/解码器/解析器/执行器/生成器/变换器...

信息显示/打印

(目标位置)

传输

目标格式的信息

交换

外设处理

协议与编码器/解码器体现了信息处理的一般性思维

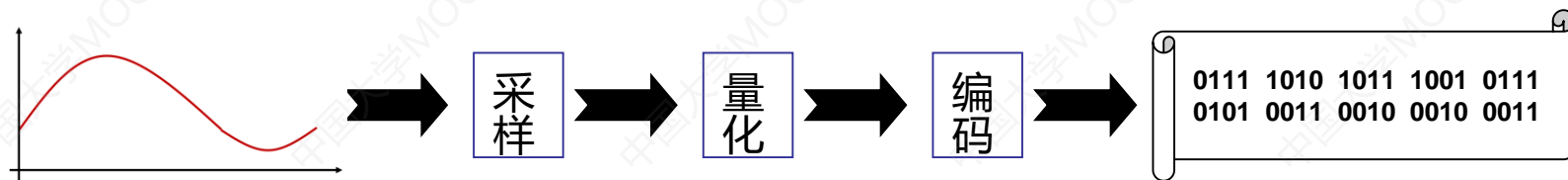
信息绑定 → 协议 → 协议编码器/协议解码器

万事万物之符号化与编码

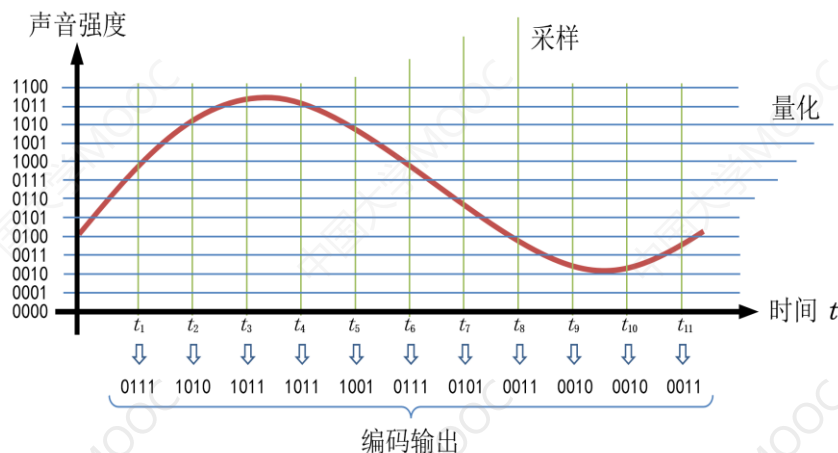
16

协议与编码器/解码器示例----音频处理

音频编码：音频是连续的模拟信号，需经采样、量化和编码后形成数字音频后，进行数字处理。所谓采样是指按一定的采样频率对连续音频信号做时间上的离散化,即对连续信号隔一定周期获取一个信号点的过程。而量化是将所采集的信号点的数值区分成不同位数的离散数值的过程。而编码则是将采集到的离散时间点的信号的离散数值按一定规则编码存储的过程。



采样频率、采样精度、编码方法及其保真度



行程编码(又称游程编码)是一种无损压缩编码，主要思路是将一个相同值的连续串用一个代表值和串长来代替。例如，字符串“SSSBBBBBMMMMNNNNNNN”，行程编码为“S3B5M4N6”。

增量编码(又称差分编码)指的是对数字数据流，除第一个元素外，将其中各元素都表示为各该元素与其前一元素的差的编码。例如，一组数据“**128**, 131, 156, 141, 121”，编码为“**128**, 3, 25, -15, -20”。

仅仅思想！

万事万物之符号化与编码

17

协议与编码器/解码器示例----图像处理

位图图像：将图像划分成均匀的由单元点构成的网格，每个单元点称为像素。每个像素可由1位或多位表示，1位只能表示黑白图像，8位能表示灰度图像，24位则能表示彩色图像。单位尺寸内的像素数目被称为图像的分辨率，由水平像素数目×垂直像素数目来表示。

*.bmp

位图文件头 BITMAPFILEHEADER bmfh

0000h 2 bytes 位图类型标识"BM"

0002h 1 dword 文件大小

000Ah 1 dword 从文件开始到位图数据开始之间的偏移量

位图信息头 BITMAPINFOHEADER bmih

000Eh 1 dword 位图信息头的长度

0012h 1 dword 位图的宽度，以像素为单位

0016h 1 dword 位图的高度，以像素为单位

001Ch 1 word 每个像素的位数

001Eh 1 dword 压缩说明（值为0则不压缩。

有其他压缩方式）

彩色表 RGBQUAD aColors[]

图像数据阵列字节 BYTE aBitmapBits[]

水平像素点数

垂直像素点数

像素点的位数

采样

量化

编码

解码

黑白-1位(0,1)

256级灰度-8位(0-255)

256色彩色-8位(0-255)

24位真彩色-24位

(红0-255、绿0-255、蓝0-255三元色)

像素表达颜色的不同，
需要编码的位数不同。

万事万物之符号化与编码

18

协议与编码器/解码器示例----图像处理

图像编码：由于位图图像的存储量大(水平像素数目×垂直像素数目×每像素位数)，通常都需要进行压缩存储，不同的压缩采用了不同的图像编码。典型的有：

◆ **JPEG**：国际标准化组织(ISO)和国际电报电话咨询委员会(CCITT)联合成立的“联合照片专家组”于1991年3月提出了JPEG标准(Joint Photographic Experts Group)。

◆ 其他常用编码格式有：**GIF, PNG, TIFF**, ...

采样

量化

编码

解码

图像
压缩

无损
压缩

有损
压缩



万事万物之符号化与编码

19

协议与编码器/解码器示例----视频处理

视频：是时间序列的动态图像(如25帧/秒)，也是连续的模拟信号，需要经过采样、量化和编码形成数字视频，保存和处理。同时，视频还可能由视频、音频及文字经同步后形成的。因此视频处理相当于按照时间序列处理图像、声音和文字及其同步问题。

视频编码：MPEG是Moving Pictures Experts Group(动态图象专家组)的缩写。提出了四个版本：MPEG-I(VCD: Video CD)、MPEG-II (DVD:Digital Versatile Disk)、MPEG-III、MPEG-IV(多媒体)。其他：国际电联H.26x系列标准，微软WMV，RM格式、FLV格式。MIDI 音乐(Musical Instrument digital Interface)，WAV，MP3等是音频的编码标准。



