目 录

张庆国

第十章　情报信息的存贮和输入输出

在前面几章中，我们从逻辑上的角度讨论了情报检索的原理。本章我们将从物理的或硬件的角度讨论有关情报信息的存贮和输入输出问题。

10.1　数据标识的代码化

无论是文献检索中的标引词、作者名、出版年，数据检索中的数据记录，还是事实检索中的知识元，都是以代码化的形式进入系统的。因为只有代码化的标识符号才能为计算机所存贮和操纵。

目前通常使用的字符编码方法有6位二进制编码的十进制代码（BCD），8位扩充的二进制编码的十进制交换代码（EBCDIC）和8位美国信息交换标准代码（ASCII）。表（10.1）和（10.2）分别给出了BCD字符代码和ASCII字符代码。

6位代码在使用时是7位的，8位代码在使用时是9位的。这是因为有1位校验位和代码本身一起使用。校验位的设置是为了能在一定程度上检查和发现数据输入时发生的技术性错误。常用的校验方法是奇偶校验。即根据该字符代码表示中“1”的个数的奇偶性来设置校验位。例如，字符“X”在BCD中的代码表示为001011，有三个“1”出现，为奇数。因此设校验位为1；再例如，字符“C”在ASCII中的代码表示为10100011，有四个“1”出现，为偶数。因此置校验位为0。

10.2　数据库的存贮载体

10.2.1　磁带数据库

1。磁带的物理特性

典型的磁带是½英寸宽的塑料带，2400英尺一盘。直径约为10½英寸，每盘重约4磅。在塑料带上镀上一层磁层。沿着磁带的纵向分成若干条磁道。其中一条磁道是用来存贮校验位的。其余的磁道根据特定计算机的编码规则用来存贮数据的字符。因此，7磁道的磁带需要与字符以6位代码编码的计算机一起使用。9磁道的磁带要与以8位代码编码的计算机一起使用。图10。1示出了一小段7磁道磁带。上面顺序用BCD代码记录了八个字符：COMPUTER。校验位为奇偶校验。

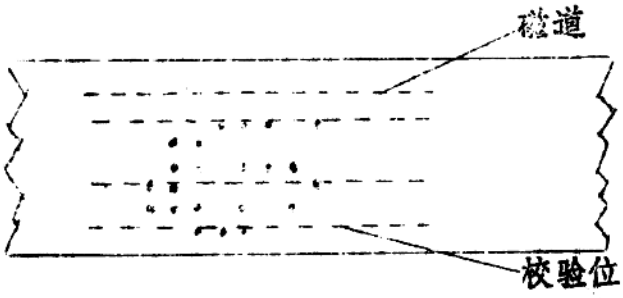

图10。1　磁带示例

一英寸长的磁带上能够存贮的字符个数称为磁道密度。测量单位为：位／英寸（BPI）。而且这些位可以存贮在每条磁带上。典型的磁道密度为800BPI或1600BPI，一盘长2400英尺。密度为800BPI的磁带，可以存贮2300万个字符。

对磁带读或写字符的速度取决于磁带驱动器的类型。每秒钟90000字符的速度是很普遍的。最多时可达每秒30万字符。由于磁带卷的惯性，所以在逐位字符之间暂停磁带是不可能的。只可以在分隔数据块的记录间隙处停止。数据块间隙的长度通常为0。6英寸。因此，在计算机内存贮器和磁带存贮器之间传送数据时，必须

整个数据块一起传送。

磁带只提供顺序存贮手段。即如果已读完磁带上的一个记录，想要读磁带上位置上的另一个记录，那么这两个记录之间的所有记录都必须先移过读写头。

2、信息的记录形式

这里要区分物理记录和逻辑记录。物理记录又称物理块，简称块。它是为了实际存贮信息和内外存之间交换信息而设置的。每个物理块是内存向外存读写数据时的最小独立单位。这就是说，一次可以读写一块或多块，但不能只读写一块的一部分。物理块的大小通常由系统硬件资源情况（内存容量，通道情况等）决定。一般为定长的（如2K），也可以是变长的。但在一次读写之前，必须由系统给定块长。

逻辑记录简称记录。它是由记录中内容所决定的。一个逻辑记录在内容上和加工处理上有独立的意义，如词典中的一个记录，数据库中的一个个人记录等。记录有时是定长的，但多数情况下是变长的，如在情报检索中。词典中的记录较容易处理为定长的，而主文档记录则很难定长。定长记录处理容易，但浪费存贮空间较大。

物理块与逻辑记录之间的关系并无硬性规定。一块中可以只有一个记录，也可以有若干个记录；一个记录可以在一块中，也可以分散在相邻的若干块中。在一块中装入几个记录后，块内还可以有空闲区域，也可以没有。当块内没有空闲区域时，显然磁带的存贮利用率要高一些，但在实际工作中，很难达到这样高的存贮利用率，尤其是对于固定长块，可变长记录时的情况。

于是，向外存读记录时，首先要调入该记录所在的物理块，然后在该块中取出该记录进行处理；向外存写记录时，给出写入位置

的外存块号和块内位移，直接将该记录写入外存中的相应位置，或者在内存中写好一块后复制到外存上。

块和记录都是读写的单位。在物理上，块可以组合成段，一段内有若干块；段又组成卷，一卷内有若干段。在逻辑上，记录组成文件，一个文件内有若干记录；文件又组成库，一个库内有若干文件。段与文件的关系，卷与库的关系，除了读写单位这一点外，与物理块和逻辑记录的关系类似。

以文献检索中的主题档磁带为例，其信息记录形式可以如下：（SM表示段标）。

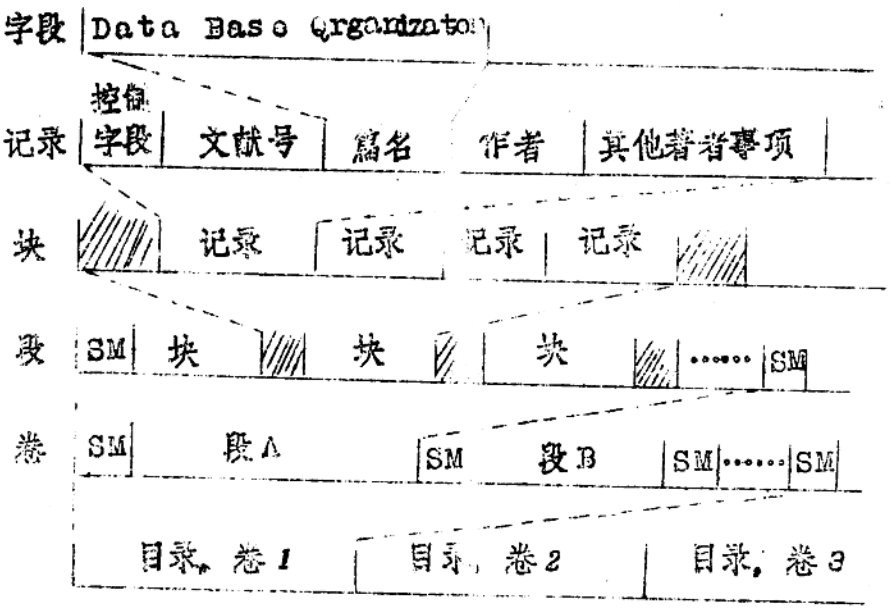

图 10.2 磁带信息的存贮形式

3。磁带数据库的用途

磁带是一种廉价的存贮介质。如果不考虑空间，在磁带上存贮数据的费用比卡片、磁盘、磁鼓以及磁芯存贮器分别低 0.20—

—60，30000和15000—600000倍。因此，对于大规模的文献数据库来说，用磁带作为存贮介质是比较合适的。

除了价格低廉外，磁带还有携带方便，复制容易等特点。因此，磁带数据库可作为商品流通。一个检索系统可购买商品磁带作为其数据库数据来源。也可以将其数据库中的文件录制在磁带上出售。目前市场上流通的文献数据库基本上都是磁带数据库。我国目前，从国外引进了多种磁带数据库，用以建立我们的检索系统。

但是，磁带数据库在费用上的节省是以其大量的存取时间作为代价换来的。磁带数据库只能顺序存取。而在情报检索（无论是文献检索、数据检索还是事实检索）中，系统需要经常随机地访问分散于数据库各个位置上的记录。因此，如果完全用磁带来建立检索用的数据库，那么频繁的倒带将耗费大量的机器时间和用户时间。这不仅是困难的，而且对于联机检索几乎是不可能的。所以，磁带数据库一般用于脱机数据库。实际联机时，将磁带上的数据库转录到磁盘等随机存取设备上，由后者完成检索任务。

还需要说明的是：商品磁带数据库中的记录格式与代码形式等往往同本系统不尽相同，因此，在转录时，一般要经过一个转换工作，有些不适合本系统的记录也可不必收入。

10.2.2 磁盘数据库

磁盘是一种随机存取的存贮器。每个磁盘上有若干条同一圆心的圆形磁道，整个磁盘又沿径向划分为若干扇区，数据记录便存放在磁道上。每个磁盘有一个读写头，读写头可沿磁盘径向移动，磁盘则绕圆心转动。存取时，首先移动读写头到所需的磁道上，然后等待所需的扇区转至读写头下面，寻到所需的位置后，磁盘继续转

动，读写头则开始工作，读入磁道上的数据，或把计算机内的数据写入磁道。图 10.3 示出一个磁盘的平面图。

在大多数情况下，磁盘是成组使用的。若干个磁盘有同一个轴连接各圆心，各磁盘同时转动。每个磁盘有一个读写头（如磁盘是双面存贮信息的，则每个磁盘有上下两个读写头）。读写头的移动时间是磁盘寻址定位时间中最主要的。为减少这一时间，一般将各磁盘的读写头保持在同一垂线上。各个磁盘的相同位置上的磁道构成一"柱面"。相应地，磁盘存贮器中的地址单位从大到小顺序是：磁盘组、柱面、扇区、块号。图 10.4 给出一个磁盘组和各磁盘的读写头。

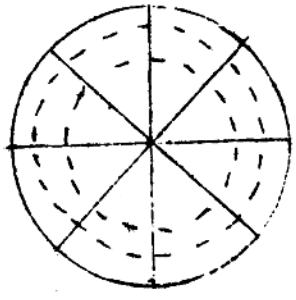

图 10.3 磁盘平面

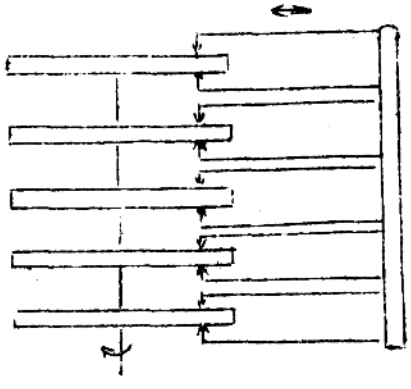

图 10.4 磁盘组

典型的磁盘组性能指标如下：

盘片数：10—50

柱面数（每盘片的磁道数）：200—400

道容量：16000 字节（16K）

总容量：1 亿字节（20 片×400 柱面×16000 字节）

旋转速度：20ms／周

最大定位时间：55ms

数据传输速率：312千字节／秒（312KBPS）

由以上性能指标可以看出，磁盘具有很大的脱机容量，而且传输速度快。对于可换磁片的磁盘组来说，价格也是不高的。更重要的是，磁盘具有很好的随机存取性，因而现在已成为联机文献情报检索系统和数据库管理系统必不可缺的外存贮器。

上面介绍的磁盘通称为硬盘。硬盘的基片一般由铝片制作的，与之相对的还有近年来发展起来的以塑料为基片的软盘。软盘的直径较小（8英寸或5英寸），容量也较小，一般为几百K字节。软盘一般可用作小型计算机系统的外存贮器。

10.2.3 其他存贮设备

1．磁鼓

磁鼓是一种鼓形存贮器，数据记录在鼓的周柱面上，有纵向的一排磁头对准鼓面上的各磁道，磁鼓旋转时，便可读出或写入磁道上的信息。

典型的磁鼓性能如下：

道数：256

道容量：32000字节

旋转速度：10ms／周

磁鼓的存贮容量是比较大的，存取时间也较快，但磁鼓存贮的费用比磁盘高得多，因此，磁鼓存贮一般不宜作为情报检索中的存贮介质。

2．卡片和纸带

卡片和纸带上记录信息后，同样可以保存下来而作为存贮文件，

但由于它们的体积大、存贮密度低、数据传输速度又很慢。故在现代计算机系统中通常只把它们作为输入输出的手段使用。我们将在以下几节中介绍它们。

10.3 情报资料的输入手段

情报资料的输入就是情报资料如何送进计算机存贮。当然，这里指的是印刷形式的情报资料，而不是磁带这样的数据记录形式。后者并不存在我们这里所指意义上的输入问题。

输入方式可以分为以下几种

1. 键到低介质方式：穿孔纸带、穿孔卡片。

2. 键到磁介质方式：磁带、磁盘等。

3. 全自动字符识别方式：光学字符识别装置（OCR）、光学标记读出装置（OMR）、磁性墨水文字识别装置（MICR）。

4. 联机终端输入方式：键盘／打印终端、无人机对话的终端、屏幕显示／键盘。

下面分别介绍

10.3.1 键到纸介质方式

这是传统的输入方式。从计算机问世以来就采用这种方法进行输入。至今仍在输入手段中占有一定地位。

纸带或卡片穿孔在穿孔机上进行人工按键，穿孔机自动将按键动作转换成与该键对应的若干凿孔动作。在纸带或卡片上凿出孔来。纸带通常有五单位纸带和八单位纸带。五单位纸带每横排有六个孔，包括一个中导孔；八单位纸带每横排有九个孔，包括一个中导孔。中导孔又称同步孔。纸带的形状类似于图10.1所示的磁带。卡片

的示样见图 10。5。每张卡片垂直方向有 80 行，水平方向有 12 列。每一行表示一个字符。每张卡片能表示 80 个字符。

纸带或卡片向计算机的输入采取光电或电刷方式。

键到纸介质方式避免了人工操作对计算机运行时间的大量延迟。从纸带或卡片向计算机输入信息的速度已比较快。例如，卡片输入机的输入速度可达 2000 张/分。但是，这样的速度相对于计算机通道的能力来说，还是非常慢的。如何提高人工打孔穿孔的速度和计算机对纸带、卡片的读入速度，仍是现在正在致力的课题。

10.3.2 键到磁介质方式

主要包括键到磁带（Key to Tape）和键到磁盘（Key to Disk），这实际上仍是一种中间存贮载体输入方式。因为这里的磁带或磁盘并非数据库，它上面的记录在正式进入数据库时还要进行处理。

键到磁介质输入方式的过程是：操作人员按键，将数据送入小型计算机的内存，计算机进行自动校验后，再将数据转录到磁带或磁盘上。

键到磁介质方式有以下几个优点：

① 可以使键盘操作者修改发觉的错误字符。因为计算机上的字符插入、删除、修改及各种编辑功能都是很容易实现的，而低介质及其设备则不具备此优点。

② 机器可以帮助校验。例如，验证校验位；若干人员重复输入。

③ 磁介质上的数据读入计算机的速度远远高于纸带输入和卡片输入。一般磁带的输入速度是每秒 20000 字符至 15000 字符，

磁盘的输入速度是每秒300000字符，并且磁介质可反复使用，也节省空间。

10.3.2 联机终端输入方式

这是主要的输入方式，主要有：

键盘／打印终端：即使用一般的电传打字机，从终端直接输入（这是用户利用分时处理的计算机）。这种方法占用计算机的时间较长，效率也不高，而且内存需要给出缓冲区。

无人机对话的终端：只有接收，没有输出，如零售输入机，数据采集设备等。

屏幕显示／键盘：是一种人机对话，兼有输入和输出两种功能的设备，联机时的操作员一般都是通过此种设备和计算机交换信息。

键到低介质，键到磁介质和联机终端输入都属于键盘输入方式。图10.6表示美国各种键盘输入方式的发展变化情况，从中可以看出，从七十年代中期开始，键到磁介质方式和联机终端输入方式的使用增长很快。

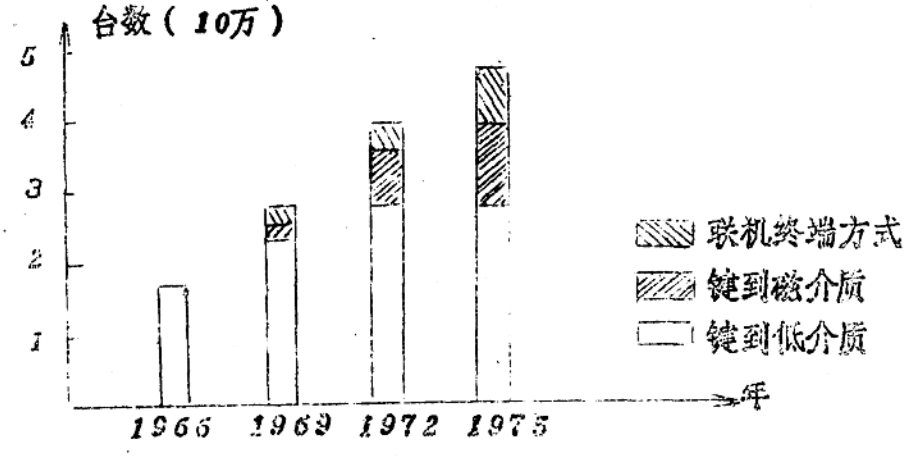

图10.6 美国各种键盘输入方式的使用情况

10.3.4 全自动字符识别方式

键盘输入是一种1对1的字符输入方式。虽然准确度很高，但它并不是完全意义上的自动数据输入方式。因为它仍部需要逐一字符按键。机器只不过是把按下的键转换成相应的代码并记录在存介质上。完全自动的数据输入方式应该取消人的干预。由机器对印刷形式的数据进行自动识别并输入。这类方法中目前使用的主要有光学字符识别装置。光学标记阅读装置和磁性墨水识别装置。

10.3.4.1 光学字符识别法

又称OCR。Optical Character Reader。一个光学字符识别装置由以下部分组成：

① 光电转换扫描部分。对输入的字符图象进行扫描。将其分解成象点。

② 图象予处理部分。将象点信息二值化（对于彩色字符图象的象点则是多值）。进行象点细化工作。并消除一部分噪声。

③ 特征抽取部分。把一个字符的象点数据根据一个特征抽取算法组成一个特征向量。待匹配判决之用。

④ 识别判决部分。将字符的特征向量经过一分类器将其分入一特定字符类内。输出。于是完成了字符识别工作。

显然。在上述过程中。特征抽取部分和识别判决部分是最重要的。同时这二者又紧密相关。特征抽取算法在很大程度上决定了识别判决函数。我们统称这两部分为识别判决部分。目前常用的识别判决方法主要有三类：第一类是根据字符的直观形象抽取特征后用相关匹配法识别。第二类是根据字符的统计特性用概率准则识别。第三类是根据字符结构用有限状态文法的句法结构识别。我们在这