计算机书目文献管理数据库

安树兰　陆玉英　孙蓓欣　编著

清华大学出版社

计算机书目文献管理数据库

安树兰　陆玉英　孙蓓欣　编著

清华大学出版社

内 容 简 介

本书系统地介绍了文献书目数据库的设计方法。全书依照数据库设计原理，分别阐述了书目数据库的几个设计阶段，特别是每个阶段的任务、利用的方法、达到的目标，并在此基础上给出了书目数据库的部分及全局模型。

全书共分四大部分。第一部分是需求调查（第一章），第二部分是设计的几个阶段（第二至第四章），第三部分是市售磁带库的特点（第五章）。为便于读者的实际设计工作，第四部分集中了大量设计参考资料（附录）。

本书可作为文献管理工作的业务人员、数据库设计人员以及计算机软件工作者的设计参考用书，也可作为大专院校图书情报及编辑专业的教材或教学参考书。对其它各种从事资料、档案工作的人员也有参考意义。

计算机书目文献管理数据库

安树兰　陆玉英　孙蓓欣　编著

责任编辑　贾仲良

☆

清华大学出版社出版

北京　清华园

北京京辉印刷厂印刷

新华书店北京发行所发行

☆

开本：787×1092 1/16　印张：23 3/4　字数：608千字

1988年11月第1版　1988年11月第1次印刷

印数：00001—10000　　定价：5.00元

ISBN 7-302-00245-2/TP·98（课）

前　言

文献的收藏、管理、宣传报道与提供工作已经有了近百年的历史，并且已经发展成一种具有相当规模的比较完整的系统，是人类社会活动中不可缺少的一个重要组成部分。然而，计算机开始进入这一行业，只有近二十年的历史。特别是在我国，研究开发工作也仅仅只有几年的历史。但是，在这不长的时期中已经清楚地看到，计算机的介入对文献工作带来了巨大的影响，并且已经促使其发生了深刻的变革。毫无疑问，随着计算机技术的进步和人们认识的加深，将把文献工作推向一个新的水平。

计算机在文献工作中的应用一般可以分成两大类：一类是情报检索，一类是文献载体的管理。前一类指的是将情报存储在机器内，按人们的检索目的，利用各种途径，查找出所需要的情报的集合。后一类则是帮助人们去管理各种类型的载有文献的载体介质，其中包括图书、期刊、微缩资料、地图、乐谱等等，管理的方面则包括采购、交换、财会统计、财产登记、编制卡片、书本目录、专题与联合目录、借还手续办理、预约等，当然也包括目录查询。这两大类应用工作有相同方面，也有不同方面。本书的研究对象是后一种类型，所以称之为书目文献数据库，也可以称之为文献管理数据库（本书中就同时使用了这两种叫法）。

一个数据库的建设往往需要三个方面的条件，一个是硬件，一个是软件，一个是数据库的设计。对于前二者的需要是明确的，当然，这些硬件和软件一定是适合要求的。对于第三方面的需要，往往被人们所忽视，或者与第二方面混为一谈。实际上二者绝不相同。数据库的设计指的是，应如何对客观世界进行调查，并将所得到的资料进行组织，按一定形式加以组配，建立相互的关系，然后再把这些数据结构，内部关系以及人们的需求科学地映象到机器内部。这种工作在建库以前是必要的，在建库之后仍然是需要的，而且一定要经过若干次往复性的循环，才能最后完成。本书着重叙述了在这一过程中将遇到的主要问题及其解决的方法，同时也用了一定的篇幅来介绍一些软件的组织方法。为了便于读者进行设计工作，在书后还附有大量的在设计书目文献数据库时需要使用的国际、国内标准和相关资料。

本书以系统设计人员为主要对象，使他们能在硬件、软件知识的基础上，结合专业知识，组织、设计一个所需要的数据库。考虑到我国目前还没有许多高级系统设计人员，还考虑到许多从事文献工作的人员或是从事计算机工作的人员也需要这些方面的知识，所以将书中的内容在范围上加以扩大，即为文献工作人员增加了一些计算机的基本知识和为计算机工作人员增加了一些必要的文献工作方面的知识。这部分内容主要反映在预篇中。

由于文献管理工作非常复杂，牵涉的内容很多，所以按一种思路来组织一部书是比较困难的，有的问题不能不前后交叉。为了全书的结构更合理一些，本书的基本思路是：先介绍调查研究的方法（第一章），再介绍数据的组织以及建立数学模型（第二章），然后介绍计算机内部实现的一些方法（第三章），还介绍了一些特殊的处理方法——非手工的传统方法（第四章）和一些可以利用的数据源（第五章）。凡是在前面的章节中利用了后面章节的内容时，我们都做了注释。对于第五章内容比较熟悉的同志，可以按一、二、三、四章的顺序读下去。对于第五章内容不太熟悉的同志，可以在读完第一章之后，先读第五章，然后再读

二、三、四章。当然也可以顺序读，把不熟悉的问题先留下来，到读完第五章之后就会解决了。

在本书的编写过程中，我们得到了很多同志的指导和帮助，特别是清华大学计算机系俞盘祥同志和中国科学院数据库筹备处沈迪飞同志，他们不单对本书的内容进行了全面的校核，更重要的是对一些重要概念进行了审核。在附录的编辑过程中，我们得到了北京图书馆张汝恭同志的大力协助。在此我们表示深深的感谢。

本书第一章由陆玉英同志编写，沈迪飞同志校核，第二、三章由安树兰同志编写，俞盘祥同志校核，第四章由孙蓓欣同志编写，安树兰同志校核，第五章由安树兰同志编写，沈迪飞同志校核。附录部分由陆玉英、安树兰二人编辑。

由于作者的知识水平与能力有限，书中难免有各种错误，恳切希望读者给予批评指正。

安树兰　陆玉英　孙蓓欣　于北京

目　录

预　篇

从第一台电子计算机问世到现在，已经有三十多年了。在这段历史中，电子计算机的发展速度在科学史上是罕见的——不仅仅它本身的数量与质量经历了巨大的变化；而且它在各个科学部门、工业企业等各行各业的广泛使用，也促进了这些部门的技术发展，使之改变了原有的面貌。从组成计算机本身的器件来看，它已经从第一代电子管计算机，第二代晶体管计算机，第三代集成电路计算机到第四代大规模集成电路计算机，目前正在向超大规模集成电路计算机发展。从它的运算速度来看，已从最初每秒几十次的计算，发展到现在的每秒钟几亿次的运算，从其体积与性能来看，其体积急剧缩小，可靠性稳步提高，同时价格也在逐年下降。特别是各种类型的现代计算机，在优良的硬件基础上，还配备了丰富的系统软件和应用软件包，使得计算机的应用范围大大扩充，并且操作难度也迅速下降。目前，它的应用已经普及到生产、科学研究、教学以及人们生活的各个方面。

在全面讨论如何利用计算机这一先进设备建立书目文献数据库之前，我们先就计算机的配置、系统软件资源、设计系统的方法以及文献工作的特点等几个方面，作简要的介绍。

§ 预1　计算机的硬件配置

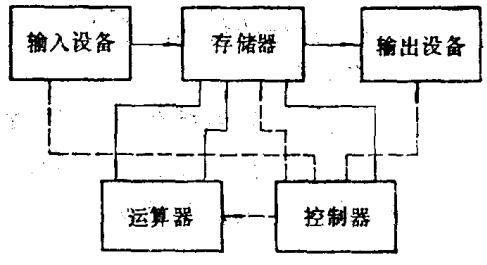
计算机的硬件配置可以分为四大类：

1．存储器——这是指内存储器，通常称之为内存。内存中有许多存储单元，在这些单元内可以存放指令，亦可以存放数据。为了识别每一个单元，必须对其进行编址，现在计算机均是按字节进行编址的。每一个地址均对应一个唯一的存储单元。

2．控制器——控制器的主要任务是指挥全机执行各种指令。它一般又分为指令部件、拍节发生器和微操作信号发生器三个部分。

3．运算器——运算器是进行四则运算和逻辑运算的部件。

4．输入/输出设备——一般又称之为外部设备。外部设备的类型很多，有些设备是输入型的，如卡片输入机，软盘输入机等，它们负责将数据输入到机器内部，另外一些设备是输出型的，如打印机等。它们是将运算的结果输出在硬拷贝(如纸张)上面的。还有一些设备则是输入/输出型的，即机器既可以从它们那里获取数据，同时，又可将数据输出到它们那里，记载下来。如磁带机、磁盘机、带显示器的终端机等等，就属于这类设备。



预 1 计算机工作原理图

计算机有了这四类部件之后，就可以工作了。其工作原理如图预1所示。

图中实线表示的是数据传送，虚线表示的是控制信号。

§ 预2 计算机的系统软件

计算机的软件是在硬件之后发展的。特别是如操作系统这一类软件的发展，是六十年代的事情。最初人们配备这类软件的目的是为了协调与控制大量的能够并发执行的程序段，同时又希望能够动态地管理各种设备，这些功能都是由操作系统提供的。后来机器上为了满足用户的要求，又增加了外部设备（主要是磁带机和磁盘机）上文件的管理。还提供了一些常用的程序，称之为实用程序。从七十年代以后，由于数据管理的应用范围大大扩充，除了可以进行各种科学计算以外，又提供了优良的数据库管理系统。这些内容都包括在系统软件的范围以内。下面将主要对操作系统部分的功能加以介绍。

操作系统的功能主要分为五个部分：

1. 处理机管理（又称之为主机管理）

它实现了机器内可以同时存在若干个作业运行。一般说来，处理机的数目总是少于作业的数目。在大多数情况下，只有一台处理机。因而，就必须要对处理机进行管理。即在不同的时刻，将它分配给不同的用户，执行不同的作业，达到多个用户共享一台处理机的目的。

2. 存储管理

存储管理包括有三个部分：内存的分配即负责将同时进行的若干作业分配在不同的存储区，以免发生冲突；内存的保护即保证各个轮流执行的作业的内存状态的正确性，同时保证一旦硬件发生故障，要有内存状态的记录，以便恢复；内存扩充部分则是为了解决作业对内存的要求大于实际可以提供的内存时出现的矛盾。

3. 作业管理

作业管理的内容包括有作业的组织，作业运行的控制和作业的调度。

用户对自己作业的控制是借助于操作系统来完成的。操作系统中的各种语言和命令提供给用户使用，以完成这一工作。它们分别是：

- 联机命令语言
- 文件管理命令和编辑命令
- 编译和执行程序命令
- 资源申请命令
- 操作命令

有关这一部分的使用，用户可以参考各种机器提供的操作手册。

4. 设备管理

广义的设备管理包括有对主机的管理，因为主机也可以看成是一种设备资源。狭义的设备管理则是指除主机以外的其它所有设备资源的管理。管理的内容有，该设备的当前状态访问，对该设备的分配以及释放。

5. 文件管理

文件是一个具有符号名的一维连续的字符序列。它们可以建立在各种类型的存储介质上。如磁带、磁盘、打印机、终端等等。文件管理是负责存取和管理文件信息的软件机构。它本身也是由目录和一组文件构成的。文件管理为用户提供了简便的、统一的存取和管理信息的方法，并建立了用户的逻辑概念和相应设备的物理文件之间的映象关系。实现了二者之间的

转换。其特点是，使用简便、安全可靠、既能提供共享，又能实现保密的要求。

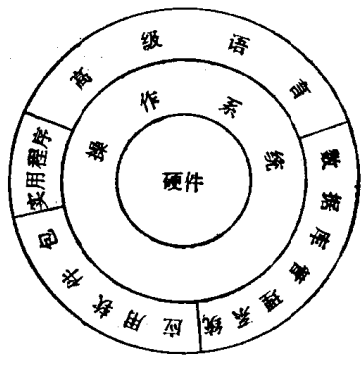
现在的计算机内，所有的信息都是以文件形式存放的，这些文件构成了一个庞大的文件库。通常，人们将其分为两大类，也就是系统文件库和用户文件库。系统文件库内装有计算机厂家提供的各种系统程序，如操作系统、实用程序、数据库管理系统等等，它们大多是以目标代码的形式给出的。除此以外，还有系统程序库，里面装有各种系统命令。它们均以独立的模块给出，这里面的程序模块也允许用户使用，也允许用户将其装配到自己的程序中去。系统文件中还有一部分则是系统参数等有关内容。

用户文件库可以有多个，每个用户都可以建立自己的用户文件库。当然，用户也可以将自己的部分文件装入系统文件库中。用户文件库中的文件由用户自己定义，可将源程序装入源程序库中，目标程序装入目标程序库中。还可以把一些特定的功能模块装入半目标程序库中，以备使用。除此以外，用户还可以为自己建立大量的数据文件，这些文件可以建立在各种外部设备上，用户可以对其规定保密性，或是为本用户专用，或是部分或全部与其它用户共享。

为了使用的方便，用户在定义文件时，可以静态地把文件分解成若干个逻辑记录。逻辑记录必须是由彼此相关的一组相邻的文件元素所构成，它也就是最小的逻辑存取单位。逻辑记录的长度可以是固定的——称之为定长记录，也可以是变化的——称之为变长记录。系统将根据记录的组织与存取方式上的不同，将其分为不同类型的文件。

文件管理系统给用户在管理数据方面提供了极大的便利条件。使得用户感到使用数据简便、安全、可靠。

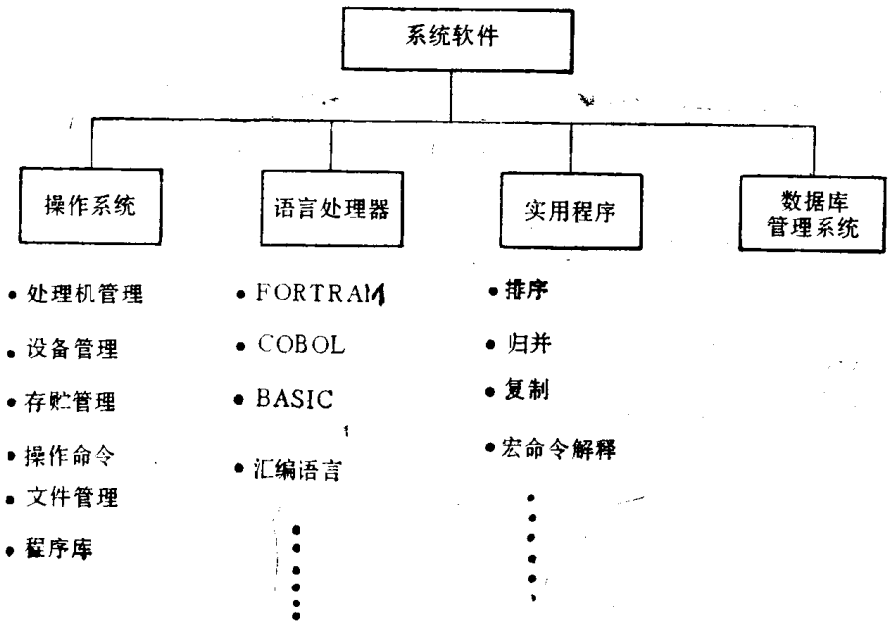
当一台机器提供了操作系统之后，用户就不再是面向硬件，而是面向操作系统进行工作了（见图预2）。



预 2 用户面前的计算机系统

这时，用户的感觉是：

• 机器具有分时性：即可以有多个用户同时使用一



预 3 系统软件

台机器，每个用户都可以单独与机器对话，仿佛是自己在独占一台机器一样。

·机器具有实时性。一旦提交了作业，就会很快得到响应。

·具有安全性：每个用户的文件可以得到很好的保护。在必要时，可以申请到不同程度的保护。比如不可为其它人阅读；可为其它人阅读，但不得为其它人修改；只可以由若干人阅读或修改。这种感觉就像是自己独占了一台机器一样。

除了这些以外，还有实用程序，高级语言以及数据库管理系统等等，就不再一一详细介绍。这些软件为人们使用机器带来极大的便利。读者可参考前页图（预3）。

§ 预3 用户文件类型

计算机为用户提供有不同类型的文件，供用户按其目的挑选使用。我们先介绍一些基本概念，然后再介绍机器上常提供使用的几种文件类型。

预3.1 逻辑记录与物理记录

计算机内存与其输入/输出设备（或外存储器）之间进行数据交换的基本单位是“块”。块有时也被叫做物理记录，即不同设备之间数据交换的基本单位是块（是物理记录）。物理记录的概念不同于逻辑记录。逻辑记录是指一组逻辑相关的数据，它将在存储器上占有一组邻接的单元。

块的大小是固定的。对于一些输入/输出设备，其块的大小是系统事先设定好的。比如读卡机是80个字符。宽行打印机是132个字符。显示终端是80个字符。另外一些存储设备的块是由用户根据个人的要求而指定的。

逻辑记录与物理记录之间有以下几种对应方式：

·一个逻辑记录占用一个物理记录，这种情况称之为不组块记录。

·一个物理记录内含有若干个逻辑记录，称之为组块记录。

·一个逻辑记录占有几个块，称之为跨块记录。

当需要在磁带、磁盘一类存储器上建立文件时，就需要对块和逻辑记录之间的关系进行选择。块越大，块内包含的逻辑记录越多，则每次内、外存交换时需要的缓冲区越大，一次交换的信息量多；块分得小，每次交换的信息量少，内存缓冲区需要的小，但是为了存取一定量的信息，内外存交换的次数会增加。所以一般情况下，当机器的规模较大，存取方式以顺序为主时，多以划分大块为好。如果存取方式以随机为主，那么就不需要划分很大的块，以节省每次交换信息的时间。

预3.2 文件的存取方法

文件的存取是指记录存入（写入）文件中或是从文件中取出（读出）记录。这种存取是通过系统来实现的。

文件是记录的集合。在本集合中，每个记录均有一个标识码。各记录按标识码的大小有一定的组织顺序，称其为逻辑顺序。这些记录在外存储器中存放有一个物理顺序。物理顺序与逻辑顺序一般是不相同的，因而必须要用一定的方法确定二者之间的对应关系，以便于对数据进行存取。即使物理顺序和逻辑顺序一致时，也必须要寻找一些方法来加快存取速度。

总之，一定要确定一个寻找数据记录地址的方法。在操作系统中的文件管理部分，就承担着这一任务。

预3.3 文件组织的基本方法

文件组织的方法基本上分为计算法和指引元法两大类。它是根据记录的使用方式、使用的频繁程度、存取的要求、存储器的性质以及容量等几种条件来选择的。一类是计算法，一类是指引元法。

1. 计算法

计算法的核心是设计一种算法，用此算法由记录的标识符计算出相应记录的地址。用户的目标是索取一个逻辑记录。

计算的方法可以有多种，基本上又可分为两类。第一类是线性计算法，第二类是杂凑（*Hash*）计算法。

线性计算法适用于顺序存储。当有N个记录时，它们按标识符的大小顺序存在外存储器上。当存储域的地址为A，每个记录的长度为m（字节）时，标识码为1，2，3，…，i，…的记录，其各记录的起始地址为

$$A,\ A+m,\ A+2m,\ \cdots,\ A+(i-1)m,\ \cdots$$

这种方法简单直观。

杂凑计算法适用的情况是标识码分布的范围广，且不连续，其核心就是要寻找一种算法，把记录标识码转换为相应记录的地址。这种算法通常是利用一个函数来实现的。该函数称之为杂凑函数（即*Hash*函数）。利用这种方法的过程是，首先确定共有多少个记录，假定该数字为N，其次就要确定存储单位的大小以及每个存储单位中可存放的记录个数。存储单位的大小，根据设备条件和存放记录长度等因素来确定。可以以一个或几个磁道为一个存储单位，或是以一个物理块为一个单位。当然，存储单位必须是可以寻址的。这种存储单位通常称之为“桶”。如果有N个记录，一个桶中可以存放n个记录，那么就至少需要$\left\lceil\frac{N}{n}\right\rceil$个桶。在这些因素确定了之后，就可以考虑利用什么函数来把记录放在相应的桶中的方法了。但是有一点要指出的是，由于记录的标识码有很大的随机性，所以必然会有这种情况发生，就是有些桶内需要装入的记录数大于桶的容量。也就是有多于n个记录标识码的映象地址是相同的。这种现象称之为碰撞。为了解决这一问题，采取了相应的解决办法，在选取容量时，要留有相当的余地，如果记录数是N，每个桶内可装的记录数是n，这时实际需要准备的桶数并不是$\left\lceil\frac{N}{n}\right\rceil$，而是比该值更大的一个数。该数为$\left\lceil\frac{N}{n}\right\rceil\times(1+10\sim20\%)$。也就是留出一些后备的桶，如若某个桶发生了碰撞现象，则可以把溢出记录放到后备桶中去。桶内记录的查找方法，可采取一些相应的方法，是比较容易的。

杂凑函数的种类比较多，例如质数除余法，基数转换法，平方取中法，折叠与移位等等。而且还有一些相应的“碰撞”处理技术。读者可以参考有关的文章。

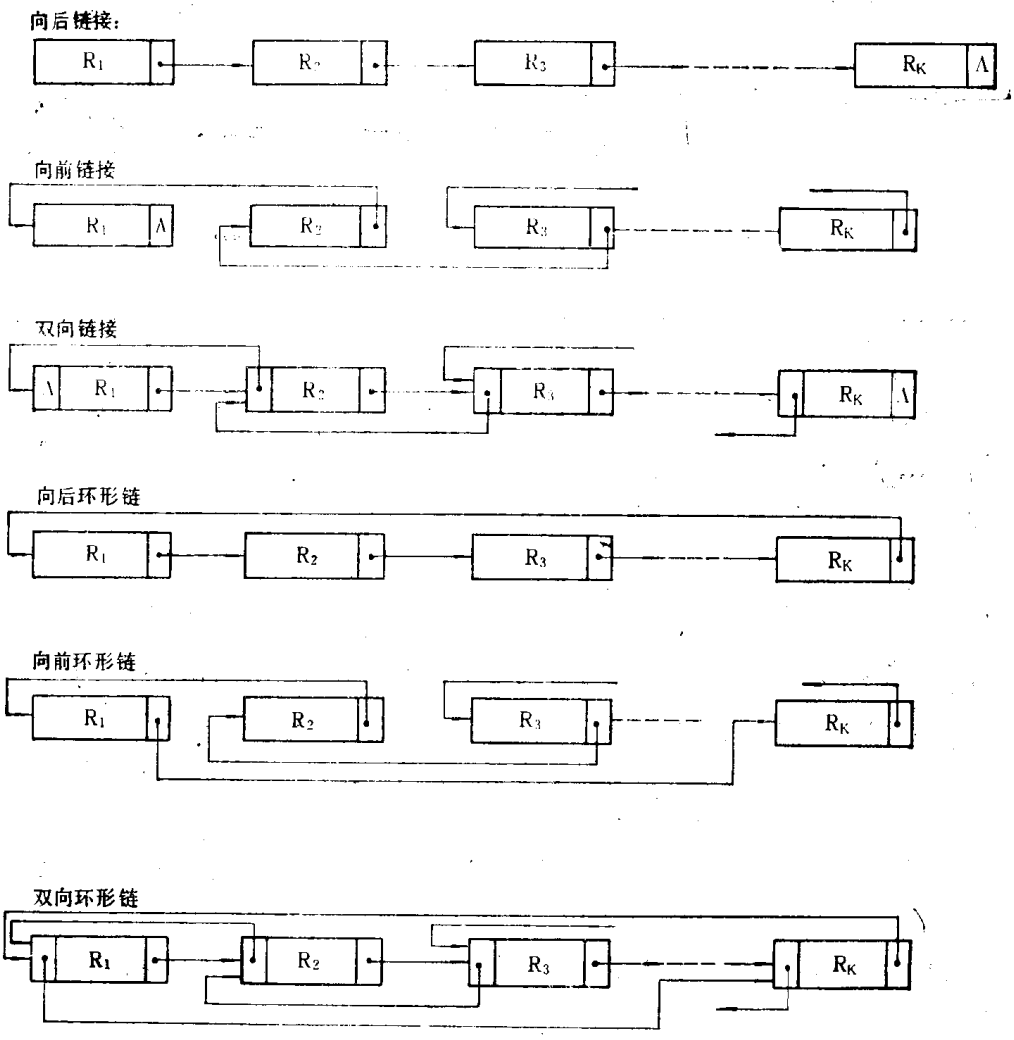
2. 指引元法

指引元法又称为指针法。就是在存储器中不单单存放数据，而且还设有一些专门的单元，存放指针，以指示各个记录的地址。这种方法就是指引方法。该方法也分为两类，第一类是索引法，第二类是链地址法。

索引法：索引法的特点是开辟一块专门的区域作为目录区，目录区中记有记录的标识码和相应于该标识码的记录地址。这个目录区就称之为索引表。索引表的每一项是相对于一条记录的，称之为该记录的索引项。

链地址法：链地址法的特点是在每个记录内嵌入这样一个数据项，它记录了与本逻辑记录相连的另一记录的地址。如果一组逻辑顺序为R_1，R_2，R_3，…，R_i，…的记录，当知道了R_1记录之后，就可以找到R_2记录……。

链地址的构成也可以有多种。可以从前向后链接，也可以从后向前链接。还可以同时建立两个链接指针项，分别指向前一个与后一个记录。链地址指针可以有链头、链尾，也可以构成一个环形链。请见图预 4 。



预 4　链地址的构成

采用比较复杂链针方法的目的，一是当记录数目很大时，可以加快存取速度。另一目的是为防止意外事故发生，某个链针被破坏时，整个链不会受完全损坏。

以上介绍的是几种基本方法。实际上各种机器上提供的文件，常常是这几种方法的组合，下面我们将重点介绍三种文件。它们是顺序文件，索引顺序文件和直接文件。

预3.4 几种常用的文件类型

1. 顺序文件

顺序文件可以建立在任何一类外部设备上。而卡片读入机、打印机、终端机以及磁带机等几类设备上，仅可以建立顺序文件。卡片读入机、打印机、终端机上的文件都是固定长记录的文件。卡片机文件的记录长度是80个字节，终端机上文件的记录长度也是80个字节，打印机文件的记录长度是80个字节或132个字节。这三类设备上的文件是不带任何标识的。

磁带机上面建立的顺序文件可以是带标识的，也可以是不带标识的。磁带以一卷为一个单位，一卷上面又可以有若干个文件。标识的作用是说明该条磁带上文件的情况。标识分为卷标识和文件标识两类。卷标识主要是说明该卷磁带的使用权，以及本卷磁带与它卷磁带的关系。文件标识是说明该文件的状况，如该文件的名称，使用权限，以及文件内的物理块数。多个用户文件可以建立在一条磁带上。同时，一个用户文件也可以建立在多条磁带上。磁带文件中的记录可以是定长的，也可以是不定长的。

磁带上面的顺序文件一般是不易更新的，特别是可变长记录的顺序文件，更是这样。

在磁盘上也可以建立顺序文件。

顺序文件的特点是物理顺序与逻辑顺序一致。在存取方式上也有很多限制，凡是建立在顺序存储器上面的文件，如磁带机、读卡机、打印机以及显示终端等等上面建立的文件，只能顺序存取。也就是要读取第i个记录，必须要先检索到第i－1个记录才可以。

建立在直接存储器上面的文件，除了可以用顺序的方法进行存取外，还可以用分块的方法或是用二分法查找。分块查找是将全部记录分为若干个组，按要查找的记录顺序码，首先确定记录在哪一个组，然后在组内再进行顺序查找即可。二分法查找是将全部记录码域一分为二，确定欲查找记录的码是落在前半部还是后半部，在确定了范围之后，再将本范围又一次一分为二，并再次确定本次所落的范围，如此继续，直至查到为止。

对顺序文件的插入、删除、修改均十分麻烦，经常要带来对文件的重新复制工作，所以不适用于经常修改文件。

由于磁带机上文件的组织方式以及各种信息格式已经达到国际统一的规格，所以磁带文件有一个最大的优点，就是可以实现不同机器之间的脱机方式的信息交换。这一点对于文献管理工作尤为重要。它为信息共享，提供了便利条件，另外一个重要的特点是，由于磁带价格便宜，信息存储量大，所以可以利用这种文件形式保留大量的工作文件副本，以防系统出现故障。一旦发生故障，可以很容易地恢复现场，而不会造成大批量信息的丢失。

2. 索引顺序文件

索引顺序文件是一种可以直接对任意一个记录进行读写的文件。由于它的读取效率高、文件易扩充，所以是广泛使用的一种文件。

索引顺序文件只能建立在磁盘上，它由两个部分组成。第一部分是索引部分，第二部分是基本顺序文件部分。

为了建立索引顺序文件，首先必须要确定该文件的主关键字。即在记录中选择一个基本项，作为主关键字。主关键字的确定原则有两个。第一，它是记录的标识项，文件中没有两个记录会具有相同的主关键字。第二，它最好是经常被用来作为查找途径的项。这是由于基本文件的物理顺序是以它的值的大小排列的。索引部分是对主关键字建立的。它记录了某个