**第十二章 文 件**

**12.1 有关文件的基本概念**

**一、文件**即**为记录的集合**，和“查找表”的差别在于，“文件”指的是存储在外存储器中的记录的集合。记录是文件中可以存取的**数据**的**基本单位**。

**二、文件**可按其中记录的类型不同而分成**两类**：

其一为操作系统的文件，文件中的记录仅是一个字符组。由于操作系统中的文件仅是一维的连续字符序列，为了用户存取和加工的方便，将文件中的信息划分为若干组，其中每一组信息称作一个记录；

其二为数据库文件，文件中的记录带有结构，是数据项的集合。记录是文件中可以存取的数据基本单位，数据项是文件中可以使用的**数据最小单位**。

**三、**记录中能识别不同记录的数据项被称为**关键字**，若该数据项能唯一识别一个记录，则称为**主关键字**，若能识别多个记录则称为**次关键字**。

**四、文件的逻辑结构**指的是呈现在用户面前的文件中记录之间的逻辑关系；**文件的物理结构**指的是文件中的逻辑记录在存储器中的组织方式。

**五、文件的操作：**

1．**检索：**

**顺序存取**：存取“当前记录的”下一个记录；

**直接存取**：存取第i个记录；

**按关键字存取**：存取其关键字等于给定值的记录。

2．**修改**：

往文件中**插入**一个或一批记录；

从文件中**删除**一个或一批记录；

**更新**文件中某个记录的属性。

3．**排序**

文件的操作方式可以**实时处理**或**批量处理**

本章讨论文件的几种常见的物理结构。

**12.2 顺序文件**

结构特点：

记录在文件中的排列顺序是由记录进入存储介质的次序决定的，即文件**物理结构中记录**的**排列顺序**和文件的**逻辑结构中记录**的**排列顺序**一致。

顺序文件的具体组织形式有两种：

**连续文件**：次序相继的两个物理记录其存储位置相邻；

**串联文件**：物理记录之间的顺序由指针相链。

操作特点：

1．**便于**进行顺序存取；

2．**不便于**进行直接存取，为取第i个记录，必须先读出前i-1个记录，对于磁盘上的等长记录的连续文件可以进行折半查找；

3．插入新的记录**只能**加在文件的末尾；

4．删除记录时，**只作标记**；

5．更新记录必须**生成新的文件**。

顺序文件的插入、删除和更新操作在多数情况下都采用**批处理方式**。此时，为处理方便，通常将顺序文件作成有序文件，称作“主文件”，同时将所有的操作作成一个“事务文件”（经过排序也成为有序文件），所谓“批处理”，就是将这两个文件“合”为一个新的主文件。具体操作相当于“归并两个有序表”，但有两点不同：（1）对于事务文件中的每个操作首先要判别其“合法性”；（2）事务文件中可能存在多个操作是对主文件中同一个记录进行的。

批处理的时间分析：

假设主文件中含有***n***个记录，事务文件中含有***m***个记录，则对事务文件进行排序的时间复杂度为O(***m***log***m***）；内部归并的时间复杂度为O(***m***+***n***），则总的内部处理的时间为O(***m***log***m***+***n***）；

假设对外存进行一次读/取为s个记录，则整个批处理过程中读/写外存的次数为

**2**×（⎡***m***/***s***⎤+⎡(***m***+***n)/s***⎤）

12.3 索引文件

**一、结构特点：**

1．索引文件由“主文件”和多级“索引”组成。

２．索引中的每个记录由“关键字”和“指针”组成。

３．通常，索引文件中的主文件是无序文件，索引是 (按关键字有序)的有序文件。

４．“索引”是在输入数据建立文件时自动生成。初建时的“索引”为无序文件，经过排序后成为有序文件。

**二、操作的特点：**

１．检索方式为：直接存取和按关键字存取。“检索”将分两步进行：先查索引，然后根据索引中指针所指索取记录。

２．插入记录时，“记录”插入在主文件的末尾，而相应的“索引项”必须插入在索引的合适位置上。因此，最好在建索引表时留有一定“空位”。

３．删除记录时，仅需删除索引表中相应的索引项即可。

４．更新记录时，应将更新后的记录插入在主文件的末尾，同时修改相应的索引项。

**三、“索引”的结构**

**１．多级静态索引**

此时的索引文件结构：

第三查找表

**⋅ ⋅ ⋅ ⋅ ⋅ ⋅**

第 二 查 找 表

**⋅ ⋅ ⋅ ⋅ ⋅ ⋅**

查 找 表

**⋅ ⋅ ⋅ ⋅ ⋅ ⋅**

索 引 表

**⋅ ⋅ ⋅ ⋅ ⋅ ⋅**

主 文 件

对主文件中**每个记录**建立一个**索引项**：

主关键字 记录在主文件中的存储位置

称作稠密索引，由这些索引项构成索引表；

从索引表建立的索引称查找表，其中每个索引项为：

最大关键字 其所在数据块的存储位置

称这类索引为非稠密索引。类似地，由查找表建立的索引为第二查找表;由第二查找表建立的索引为第三查找表。

按关键字进行检索时，从第三查找表开始，至多访问外存五次。

**２．动态索引**

索引表采用查找树表或哈希表。**优点**：

1. 不需要建立多级索引;
2. 初建索引不需要进行排序;
3. 插入或删除记录时，修改索引方便;

用查找树表作索引时，查找索引所需访问外存次数的最大值恰为查找树的深度。可以作索引的树表有：二叉排序树、B-树和键树

稠密索引的优点是，可以实现“预查找”

缺点是，索引表占用的存储空间大。

**12.4 索引顺序文件**

结构特点：

主文件按主关键字有序，对一组记录建立一个索引项(建立非稠密索引)。

有两种典型的索引顺序文件

**一、ISAM文件**

ISAM(**I**ndex **S**equential **A**ccess **M**ethod)(索引顺序存取方法)是一种专为磁盘存取设计的文件组织方法。

１．文件的组织方式：

主文件按柱面集中存放，同时建立三级索引：磁道索引、柱面索引和主索引。

**磁道索引结构**

关键字 指针 关键字 指针

**基本索引项**  **溢出索引项**

磁道索引

r(14) r(21) r(38)

r(41) r(57) r(63)

r(72) r(85) r(99)

99

柱 面 索 引

溢 出 区

一个柱面

210

•

210 •

主 索 引

•

•

•

•

•

•

1024 •

•

1024

磁道索引

r(514) •••••

••••• r(1024)

溢 出 区

２．操作的特点：

检索：可有两种方式：

顺序存取— 依关键字最小至大顺序存取

按关键字存取— 从主索引开始，到柱面索引，到磁道索引，最后取得记录，先后访问四次外存。

插入:

将记录插入在某个磁道的合适位置上;

将该磁道上关键字最大的记录移出到本柱面的溢出区中;

修改本磁道的索引项(包括基本索引项和溢出索引项)。

删除:

在被删记录当前存储位置上作“删除标记”。

３．文件重组

在经过多次的插入和删除操作之后，大量的记录进入文件的“溢出区”，而“基本存储区”中出现很多已被删去的记录空间，此时的文件结构很不合理。因此，对ISAM文件，需要周期地进行重整。

４．柱面索引的位置

ISAM文件占有多个柱面，其柱面索引应设在数据文件的中间位置上，以使“磁头”的平均移动距离最小。

**二、VSAM文件**

VSAM(**V**istual **S**torage **A**ccess **M**ethod)文件是利用操作系统中提供的虚拟存储器的功能组织的文件，免除了用户为读/写记录时直接对外存进行的操作，对用户而言，文件只有控制区间和控制区域等逻辑存储单位。

**１．文件的结构**

由索引集、顺序集和数据集三部分组成。

索引集

B+树

••• ••• 顺序集

数据集

••• ••• •••

控制区域 控制区间

数据集内含若干控制区域，而控制区域内含若干控制区间，每个控制区间内含一个或多个记录，当含多个记录时，同一控制区间内的记录按关键字自小至大有序排列，且文件中第一个控制区间中记录的关键字最小;

顺序集内存放的是数据集的索引，每个控制区间有一个索引项，它由两部分信息组成：该控制区间中记录的最大关键字和指向该控制区间的指针。若干相邻控制区间的索引项形成顺序集中的一个结点，结点之间用指针相链;

索引集是顺序集的索引，即文件的高层索引项，也由最大关键字和指针两部分信息组成。

从索引文件的角度看，数据集即为主文件，而顺序集和索引集构成“索引”。

２．**控制区间**是用户进行一次存取的逻辑单位，可看成是一个逻辑磁道。但它的实际大小和物理磁道无关。

**控制区域**由若干控制区间和它们的索引项组成，可看成是一个逻辑柱面。

VSAM文件初建时，每个控制区间内的记录数不足额定数，并且有的控制区间内的记录数为零。

３．**顺序集**本身是一个单链表，它包含文件的全部索引项，同时，顺序集中的每个结点即为B+树的叶子结点，**索引集**中的结点即为B+树的非叶结点。

４．**文件的操作**

检索：可进行顺序存取和按关键字存取;

插入：按关键字大小插入在某个适当的控制区

间中，当控制区间中的记录数超过文件

规定的大小时，要“分裂”控制区间，

必要时，还需要“分裂”控制区域;

删除：必须“真实地”删除记录，因此要在控

制区间内“移动”记录;

５．**VSAM文件**通常被作为大型索引顺序文件的标准组织方式。

其**优点是**：动态地分配和释放空间，不需要重组文件；能较快地实现对“后插入”的记录的检索;

其**缺点是**：占有较多的存储空间，一般只能保持约75%的存储空间利用率。(因此，一般情况下，极少产生需要分裂控制区域的情况)

**12.5 直接存取文件**

**１．**和前几节讨论的文件组织方法不同，**直接存取文件的特点**是，由记录的关键字“**直接”得到**记录在外存上的**映象地址**。

类似于哈希表的构造方法，根据文件中关键字的特点设计一种“哈希函数”和“处理冲突的方法”将记录散列到外存储设备上，又称“散列文件”。

**２．哈希文件的结构**

由于记录在外存上是成组存放的，因此允许多个记录映象到同一个地址上。在此，称外存储器中存放多个记录的“数据块”为“桶”。因此由哈希函数得到的映象地址为“桶地址”。例如：有一组关键字如下所列

{589，063，269，505，764，182，166，330}

假设哈希函数为 key MOD 7，每个桶可以容纳3个记录（称桶的容量为3），则可得哈希文件如下所示：

基 桶 溢 出 桶

0 063 182

1 589 505 764 330

2

3 269

4

5 166

6

在哈希文件中，**“冲突”和“溢出”是不同的概念**。一般情况下，假设桶的大小为*m*，则允许哈希地址产生m-1次的冲突，当发生第m次冲突时，才需要进行“冲突处理”，对散列文件而言，通常采用链地址法出路冲突。为区别起见，称直接“散列”的数据块为**“基桶”**，而因“溢出”存放的数据块为**“溢出桶”**。

**３．文件的操作**

检索：只能进行按关键字的查找，不能进行顺

序查找。检索时，先在基桶内进行查找，

若不存在，则再到溢出桶中进行查找。

插入：当查找不成功时，将记录插入在相应的

基桶或溢出桶内。

删除：对被删记录作特殊标记。

**４．优点：**记录随机存放，不需要进行排序；插入、删除方便，存取速度快；节省存储空间，不需要索引区。

**缺点：**不能进行顺序存取；在经过多次插入和删除操作之后，需进行“重组文件”的操作。

**12.6 多关键字文件**

**一、多关键字文件的特点**

除需要对主关键字建立“主索引”外，尚需对各个次关键字建立“次索引”。

次索引项： 次关键字 （指向记录的）指针

**二、次索引的组织方法**

1．**多重链表文件**

特点：将所有**具有相同次关键字的记录链接在同一链表中**，该链表的头指针即为次索引项中“指针域”的值。

2．**倒排文件**

特点：将所有**具有相同次关键字的记录构成一个次索引顺序表**，此时的次索引顺序表中仅存放记录的“主关键字”或记录的“物理记录号”。次索引项中的“指针”指向相应的次索引顺序表。

3．**次关键字索引表本身的结构**可以是**顺序表**，也可以是**树表**或**哈希表**，视具体的次关键字的特性而定。