**第7章 文件与磁盘排序 学习指导**

1、时间

2020年5月19日周二 （1-4班3-4节，5-10班、数学、大数据专业5-6节）

2020年5月21日周四 （1-4班3-4节，5-10班、数学、大数据专业5-6节）

2次课共4学时。

2、内容

电子教材&PPT预习；

腾讯课堂讲解本章重点和难点；

QQ群交流练习题（中间根据需要采用QQ视频或腾讯视频讲解）；

具体安排：

|  |  |  |
| --- | --- | --- |
| 1 | 5月19日 | (1) 查找算法的实现以及各种查找方法的时间性能； |
| 2 | 5月21日 | (2) 在树结构和散列结构上进行查找的基本思想和方法；  (3) 基于关键字查找与基于关键字散列地址查找的本质区别。 |

**3、知识要点**

**文件：**是性质相同的记录的集合。记录是文件中存取的基本单位，数据项是文件可使用的最小单位，数据项有时称字段或者属性。

　　·逻辑结构是一种线性结构。

·操作：检索和维护。并有实时和批量处理两种处理方式。

　　·存储结构是指文件在外存上的组织方式。

·基本的组织方式有：顺序组织、索引组织、散列组织和链组织。

·常用的文件组织方式：顺序文件、索引文件、散列文件和多关键字文件。

　　评价一个文件组织的效率，是执行文件操作所花费的时间和文件组织所需的存储空间。

　　检索功能的多寡和速度的快慢，是衡量文件操作质量的重要标志。

**顺序文件：**是指按记录进入文件的先后顺序存放、其逻辑顺序和物理顺序一致的文件。主关键字有序称顺序有序文件，否则称顺序无序文件。

　　一切存储在顺序存储器（如磁带）上的文件都只能顺序文件，只能按顺序查找法存取。

　　顺序文件的插入、删除和修改只能通过复制整个文件实现。

**索引文件：**通常是在主文件之外建立一张索引表指明逻辑记录和物理记录之间一一对应的关系，它和主文件一起构成索引文件。

　　索引非顺序文件中的索引表为稠密索引。索引顺序文件中的索引表为稀疏索引。

　　若记录很大使得索引表也很大时，可对索引表再建立索引，称为查找表。是一种静态索引。

　　索引顺序文件常用的有两种：

·ISAM索引顺序存取方法：是专为磁盘存取文件设计的，采用静态索引结构。

·VSAM虚拟存储存取方法：采用B+树作为动态索引结构，由索引集、顺序集、数据集组成。

**散列文件：**是利用散列存储方式组织的文件，亦称为直接存取文件。

·优点是：文件随机存放，记录不需要排序；插入删除方便；存取速度快；不需要索引区，节省存储空间。

·缺点是：不能进行顺序存取，只能按关键字随机存取，且询问方式限地简单询问，需要重新组织文件。

**多重表文件：**对需要查询的次关键字建立相应的索引，对相同次关键字的记录建一个链表并将链表头指针、长度、次关键字作为索引表的索引项。

**倒排表：**次关键字索引表称倒排表，主文件和倒排表构成倒排文件。

**外部排序：**大文件排序，待排序的记录存储在外存储器上，在排序过程中需要经过多次的内、外存之间的交换。外部排序可分为磁盘文件排序和磁带文件排序。其基本方法都是归并。

**外部排序的知识点组要包括：**

多路平衡归并、置换-选择排序、最佳归并树

**四、习题**

1. **综述文件有哪几种常用的组织方式，它们各有什么特点。**
2. **简单比较文件的多重表和倒排表组织方式各有什么不同。**
3. **假设某文件经内部排序得到100个初始归并段，试问:**

**(1)若要使多路归并三趟完成排序，则应取归并的路数至少为多少？**

**(2)假若操作系统要求一个程序同时可用的输入、输出文件的总数不超过13，则按多路归并至少需几趟可完成排序？如果限定这个趟数？则可取的最低路数是多少？**

1. 假设一次I/O的物理块大小为150，每次可对750个记录进行内部排序，那么，对含有150000个记录的磁盘文件进行4-路平衡归并排序时，需进行多少次I/O？
2. 试问输入文件在哪种状态下经由置换-选择排序得到的初始归并段长度最长？其最长的长度是多少？
3. **试问输入文件在哪种状态下经由置换-选择排序得到的初始归并段长度最短？其最短的长度是多少？**
4. 已知某文件经过置换-选择排序之后，得到长度分别为47, 9, 39, 18, 4, 12, 23和7的八个初始归并段。试为3-路平衡归并设计一个读写外存次数最少的归并方案，并求出读写外存的次数。
5. **已知有31个长度不等的初始归并段，其中8段长度为2，8段长度为3，7段长度为5，5段长度为12，3段长度为20（单位均为物理块）。请为此设计一个5-路最佳归并方案，并计算总的（归并所需的）读/写外存的次数。**