308-420A Secondary storage Algorithms and Data Structures

Supplementary Questions and Exercises

Sequential Files

Section 1.2–5, Sequential Files

1. A file of 3.2 million records is to be sorted using replacement-selection and merge techniques. Buffer capacity in RAM for replacement-selection is 160 records. The subsequent merges are 10-way merges. How many initial runs are expected and what is the resulting number of merge passes?

   (a) 10,000 initial runs, 4 merge passes.

   (b) 10,000 initial runs, 1000 merge passes.

   (c) 20,000 initial runs, 5 merge passes.

   (d) 20,000 initial runs, 15 merge passes.

   (e) 20,000 initial runs, 2000 merge passes.

2. Sequential files are always good for

   (a) activities $< 1\%$.

   (b) volatile files.

   (c) queries requiring high symmetry, when the file is sorted.

   (d) More than one of the above.

   (e) None of the above.

3. The following numbers are read by the replacement selection algorithm into RAM with a capacity for three numbers. How many runs would you *expect* to be generated, and how many runs actually are generated?

   22 9 12 7 6 14 24 19 16 3 17 2

4. A transaction retrieves (a) nine records from a sequential file of 9000 records with 90 records per block and (b) nine records from a direct file of 9000 records with 100 records per block. What is the activity in each case?

5. For a sequential file of N records, how many records must be read on the average for successful and unsuccessful searches a) if the file is ordered and b) if it is unordered?

6. Initial runs are being created by a replacement selection program at an average of two per second. The unsorted data is being read by the program at 1000 records per second. What is the size of the buffer used by the program?

7. Replacement selection generates a million initial runs for a certain file. The merge program can merge ten runs into one. Including the initial runs, how many times will each record in the file get written? That is, how many write passes are needed to sort the file?

8. Describe two situations in which a sequential file should be sorted. Explain why sorting is an advantage in each case. (Note. Check the next question before answering this one.)

9. Describe a situation in which a sequential file should not be sorted. Explain why sorting is a disadvantage in this case.

10. (a) Is there data for which replacement selection will generate exactly one run, no matter how much data? Which data?

    (b) Is there data for which replacement selection will generate runs whose sizes are all exactly equal to the RAM capacity? Which data?

11. The following sequences of integers are input to the replacement-selection phase of a sort for secondary storage. There is room in RAM for only two integers. What is the average run length in each case? a) `9, 1, 8, 2, 3, 7, 6, 4, 5` b) `8, 7, 6, 5, 4, 3, 2, 1`

12. a) The merge phase of a sort for secondary storage has 1,000,000 initial runs as input, and does 10-way merges. How many passes of the data will be needed to complete the sort? b) If RAM capacity during replacement selection is 10,000 bytes, how many bytes long is the sorted file, assuming the statistically expected number of initial runs were generated?

13. How should a k-way merge be terminated?

14. Searches can be *successful* or *unsuccessful*. Sequential files can be *ordered* or *unordered*. For each possible combination, write down the expected number of accesses that will be made to a file of n blocks under uniform usage for a transaction which requests two records.

15. Suppose we must do a merge sort, with replacement selection, for two million records. RAM has capacity for 1000 records. How many passes of the data are needed a) in the best case; b) in the average case; c) in the worst case? (Assume that records are read in one at a time during the merge phase.)

16. In terms of expected costs for finding a single record, what are the practical differences between ordered sequential and unordered sequential files?

17. Draw the flowchart for, or otherwise briefly describe, the 2-way merge algorithm that reads the following two files (sets of records) as input

    `1, 2, 4, 7` `2, 3, 4, 6`

    and gives the following output

    `1, 3, 6, 7`

18. Use replacement selection with a RAM capacity of three records to make initial runs for the following records, input in the order given.

    `Tom, Sue, Sam, Pat, Nan, May, Mac, Joe, Jim`

    How many initial runs are generated? Show them.

19. A file of 2 billion records (2×10^9) is to be sorted using replacement selection with RAM capacity of 1000 records, and then 100-way merges. How many *passes* of the file should be expected to do this?

20. a) Using replacement selection with a capacity of two records, generate the initial runs for the "records" (1 letter each) `T,H,E,Q,U,I,C,K,R,E,D,F,O,X`, entered in this order. b) Discuss the average run length in terms of the theoretically expected value. Explain the lengths of the shortest and longest runs.

21. Using the initial runs you got in the previous question, finish sorting the file with a 2-way merge. How many passes did this take? Compare this with the theoretically expected value.

22. If the unsorted and sorted files of the last two questions are sequential with two records per block, how many blocks will a search for "`K`" read in each case? Compare and discuss the comparisons from the perspective of theory.

23. Twenty-seven initial runs are sorted at a cost of three read/write passes of the data. How many passes would be needed if these merges were all done in RAM?

24. An unordered sequential file with n blocks and a 70-30 usage distribution of records is searched separately for each of a set of records, none of which is in the file. What is the average number of blocks accessed per search?

25. Show the flowchart for a program to "update" one file with another. For example, if file {`(a,x),(b,x),(c,x)`} is updated by file {`(b,y),(d,y)`}, the result should be the new file {`(a,x),(b,y),(c,x),(d,y)`}, where `a, b, c, d` are values of the field controlling the update, and `x, y` are values of another field in the records. (Assume both files are suitably sorted, and have no duplicates.)

26. A file of n blocks is sorted on a computer with buffer capacity of r blocks for replacement-selection. Suppose every initial run has the theoretical expected length, and that subsequent merges are all three-way. Give the formula for the number of *passes* required to sort the file.

27. For successful and unsuccessful searches for r records from ordered and unordered sequential files, what are the expected numbers of accesses needed by a good search algorithm? (The records in the files are in blocks, of course.)

	ordered	unordered
successful		
unsuccessful		

28. Show the flowchart for the merge that finds the symmetric difference between two sets (that is, elements in either set that are not in the other set).

29. How does an ordered sequential file differ from an unordered sequential file a) for single search b) for high activity?

30. Suppose you have s sorted sequential files of n blocks each. a) How many accesses would it cost for an s-way merge algorithm to find the intersection of all files? b) Describe the merge logic needed to do this (apart from stopping conditions).

31. A sequential file of 100 million 100-byte records is to be queried by another sequential file of 1 million 20-byte search "keys". Draw the flowchart needed to expand each of the search keys to full, 100-byte records.

32. For the query of the previous question, what proportion of the first file must be processed before all the keys have been found (all other things being equal)?

33. If the key file in question 5 had, in addition, a second set of keys, each 5 bytes long, which are to be looked up independently of the first set of keys, how many passes of the main file would be needed?

34. An unordered sequential file is to be searched, using merge logic, for all the records whose search values (the values of the field being searched on) are stored in another, much smaller file. a) Which operation from set theory corresponds to the search to be done? b) What is the (big-O) complexity of the entire search? (Explain your symbols.)

35. Replacement selection, with an internal capacity of 4 records, is used to construct initial ordered "runs" from the following 16 records.

    `19,17,23,25,2,14,18,16,7,5,11,10,6,22,20,21`

    a) How many runs will be produced? b) How many runs are to be expected if all we knew is that 16 records came in.

The author gratefully acknowledges support from the taxpayers of Québec and of Canada who have paid his salary and research grants while this work was developed at McGill University, and from his students (who built the implementations and investigated the data structures and algorithms) and their funding agencies.