Examiner's Report on the Ph. D thesis

Examiner: Tadao Takaoka

Title: Sequential and Parallel Algorithms for the Generalized Maximum Subarray Problem

Candidate: Sung Eun Bae

Summary: The maximum subarray problem is to find a subarray in the given array that maximizes the sum in it. The values of array elements take real numbers. The given array is one-dimensional of size n or two dimensional of size (n, n). If array elements are non-negative, a trivial solution is the whole array. Thus the mean value is subtracted from each element. Several efficient sequential algorithms are known in the literature. Application areas are in graphics and data mining. In graphics, we can identify the brightest spot, and in data mining we can identify the most romising customer range.

The thesis generalizes this problem in two directions. One is to find the maximum, second aximum, ..., K-th maximum subarray, called the K maximum subarray problem. There are two variations; one allows overlapping among the found subarrays, the other all disjoint. The thesis achieves an algorithm of $O(n^3 + K\log n)$ time for the former problem and that of $O(n^3 + n^2 K\log K)$ time for the latter. Those are algorithms for the two-dimensional problems, based on efficient one-dimensional algorithms. New data structures, called persistent heaps and persistent tournaments are used for these algorithms.

In the second part, the computational model is generalized from one processor to multiple processors for parallel processing. Specifically a mesh architecture is employed with n^2 processors. The thesis achieves O(n) time for the maximum subarray problem and O(Kn) time for the K maximum subarray problem. The problem here is a general one where overlapping is allowed. In the following, each chapter is reviewed.

Review of each chapter

Chapter 1. This chapter introduces historical backgrounds and possible application areas. In many papers, the origin of the maximum subarray problem is attributed to the two papers by Bentley, published in 1984. The candidate discovered a pioneering paper by Grenander published in 1977, which is a good contribution for historical investigation. As possible application for the one-dimensional case, genome sequence analysis is discussed, and for the two-dimensional case, applications for graphics and analysis for sales database analysis are introduced. For the latter application, a subarray of rectangular shape is good enough. For the first application in graphics, we do not need to restrict to rectangular shape. Perhaps, this issue should be addressed.

Chapter 2. This chapter provides precise problem definitions and mathematical and algorithmic concepts used in the later developments. For the preparation for the K-maximum subarray problem, a selection algorithm of linear time is introduced. This algorithm is not very efficient in practice, as compared to the FIND algorithm by Hoare, whose expected time is linear. I suppose the candidate chose the present selection algorithm, since the thesis is entirely devoted to worst case analysis. Probably the candidate should mention this.

Chapter 3. This chapter defines the overlapping K-maximum subarray problem and develops two major algorithms for this problem. The first one is based on persistent tournament trees, resulting in O(nlogK + KlogK) time. The next one is based on persistent binary heaps, resulting in O(n + Klogn) time. Although the second is simpler and more efficient in the time complexity, the first has some merits if modified to solve the length constrained maximum subarray problem. This development is a very new topic, but the candidate is welcome to make a small comment to highlight the importance of the first approach.

Chapter 4. This chapter defines the disjoint k-maximum subarray problem and develops efficient algorithms for that problem. The definition is to find k maximum subarrays successively from the remaining array portion. If we do not require sorted order, Ruzzo-Tompa's liner time algorithm is the most efficient, but this algorithm is not easily extensible to two dimensions. The candidates develops an algorithm based on a tournament tree, not persistent this time, but with a new feature of hole-manipulations. That is, already found subarrays are removed from the tree data structure as a hole. This version happens to be extensible to two dimensions well, resulting in $O(n^3 + n^2KlogK)$ time. As the two-dimensional disjoint problem was not researched before, the contribution here is high.

Only small comments here are that the problem is defined to choose maximum subarrays from the remaining portion. This leads to a situation where the algorithm may not choose K maximum subarrays in two dimensions if there is a tie. This problem needs further research, but at least the candidate should mention this problem.

Chapter 5. This chapter is devoted to computer experiments. The analysis of computing time for two dimensions for large K is not easy. The candidate explains the non-linear behavior based on the tournament tree manipulation. As the time is the sum of everything, this reasoning may be correct, or some other factors might be involved. At least the linearity in n is established and the overall trends reflect well the theoretical analysis.

Chapter 6. This chapter discusses future possibilities for sequential algorithms for the K-maximum subarray problem. Partially overlapping problems may be more practical for applications and interesting for the future research.

Chapter 7. This chapter introduces parallel computational models for the later chapters. Specifically PRAM, mesh architecture, and cube connected models are discussed. For its simplicity, mesh architecture is chosen as the model for parallel implementation of algorithms for the K-maximum subarrays.

Chapter 8. This chapter designs an (n, n)-mesh computer that computes the maximum subarray in O(n) time and K maximum subarrays in O(Kn) time. The problems of the overlapping and disjoint case cases are treated. The disjoint case can be implemented by repeated use of the maximum subarray after putting minus infinity at the found portion. The overlapping K-maximum subarray version is a slight modification of the maximum subarray version, in which K data items are sent from cell to cell. As the data amount for two dimensional problems, such as graphical images, is enormous, the sequential algorithms discussed in previous chapters may not cope with real time applications. Thus the use of simple parallel architecture can be justified for speed up.

Chapter 9. This chapter discusses some variations of the previous parallel architecture. The first is the issue of data loading. If graphic images, such as video images, come from outside one by one, the issue of data loading becomes a great concern. If the mesh computer needs to wait until the data stabilize at appropriate locations, we lose efficiency for the real time environment. The chapter discusses what is called the run-time data loading, which is a sort of pipe-lining process, and prevent the loss of computing time by data loading. The second issue is the scalability. The number of processors needed, n^2, may be too much in practice. If we have a limited number of processors, we can still speed up the computation by instructing each processor to simulate several cells. This is well explained. The third is the bidirectional approach. The architecture in Chapter 9 is based on horizontal and vertical data movement, but the movements are mono-directional, that is, from left to right, and from top to down. The candidate introduces both directions horizontally, and establishes 1.5n steps. I conjecture that 1.0n steps may be possible by introducing bidirectional approach both horizontally and vertically.

Chapter 10. This chapter summarizes mesh implementations and discusses future possibilities. The problem of broadcasting for the disjoint k maximum subarrays is challenging.

Conclusion and Recommendation

The thesis results were published in four international conference proceedings and two journals. The thesis opened two original areas: the k maximum subarray problem and the mesh architecture for the maximum subarray and K-maximum subarray problem. The first area attracted many researchers world-wide, and the candidate's publication is always cited as the starting point. The second contribution can open the theory to practical implementation that can give us the solution in realistic amount of time. The result is patented with the New Zealand Government.

Based on these observations, I recommend that the candidate be awarded the degree of Doctor of Philosophy and the thesis placed on the Dean's list. The suggestions given in the review section are not mandatory, but can be accommodated in a possible revised version in case one is made.