Streaming Algorithms for Geometric Problems

Piotr Indyk
Computer Science and Artificial Intelligence Lab
MIT

1 Introduction

Computing over data streams is a recent phenomenon that is of growing interest in many areas of computer science, including databases, computer networks and theory of algorithms. In this scenario, it is assumed that the algorithm sees the elements of the input one-by-one in arbitrary order, and needs to compute a certain function of the input. However, it does not have enough memory to store the whole input. Therefore, it must maintain a "sketch" of the data. Designing a sketching method for a given problem is a novel and exciting challenge for algorithm design.

The initial research in streaming algorithms has focused on computing simple numerical statistics of the input, like median [MP80], number of distinct elements [FM85] or frequency moments [AMS96]. More recently, the researchers showed that one can use those algorithms as subroutines to solve more complex problems; see the survey [Mut03] for detailed description of the past and recent developments. Still, the scope of algorithmic problems for which stream algorithms exist is not well understood. It is therefore of importance to identify new classes of problems that can be solved in this restricted settings.

Recently, several authors proposed streaming algorithms for geometric problems [AHP01, FKZ02, TGIK02, Ind03, CM03, MS03, HS04, BCEG04, Cha04, STZ04, Ind04, HPM04, FIS04][1]. Most of the algorithms appear to follow one of the two approaches:

1. **Merge and Reduce**. The main component of this approach is an (off-line) algorithm which, given a set of points P, constructs a subset of P which is small but nevertheless 'well-approximates' P. The algorithm is then applied in a divide-and-conquer manner to the whole data stream. Because of the tree-structured computation, the algorithm can be simulated using small memory.

   This very general approach goes back to [MP80], and has been first adapted to the geometric setting in [AHP01, AHPV], using the notion of a *core-set*. However, it does not (appear to) work in the dynamic setting, where the points can be *deleted* as well as inserted.

2. **Randomized embeddings**. This approach relies on approximate embeddings of the set of points represented by a stream into a low-dimensional space (see [Ind01] for more information on embeddings). Specifically, the set of points is mapped to a low-dimensional vector, typically using a random linear map or its variants. The vector is then used to estimate the desired statistics of the data, such as the cost of the minimum tree spanning the set of points.

   In the context of streaming algorithms, this approach goes back to [AMS96] or even [FM85]. The use of linear mappings enables to perform point deletions as well as insertions. However, the resulting algorithms have typically higher complexity than their insertions-only counterparts.

In this talk I will present an overview of the known results, and describe the aforementioned methods. I will also present a list of open problems in the area.

[1] Several other algorithms were proposed for the more general metric space setting [CCFM97, GMMO01, Mey01, COP03].

References

[AHP01] P. Agarwal and S. Har-Peled. Maintaining approximate extent measures of moving points. *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 2001.

[AHPV] P. K. Agarwal, S. Har-Peled, and K. R. Varadarajan. Approximating extent measure of points. *Journal of the ACM*.

[AMS96] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. *Proceedings of the Symposium on Theory of Computing*, pages 20–29, 1996.

[BCEG04] A. Bagchi, A. Chaudhary, D. Eppstein, and M. T. Goodrich. Deterministic sampling and range counting in geometric data streams. *Proceedings of the ACM Symposium on Computational Geometry*, 2004.

[CCFM97] M. Charikar, C. Chekuri, T. Feder, and R. Motwani. Incremental clustering and dynamic information retrieval. *Proceedings of the Symposium on Theory of Computing*, 1997.

[Cha04] T. Chan. Faster core-set constructions and data stream algorithms in fixed dimensions. *Proceedings of the ACM Symposium on Computational Geometry*, 2004.

[CM03] G. Cormode and S. Muthukrishnan. Radial histograms for spatial streams. *DIMACS Tech Report*, 2003.

[COP03] M. Charikar, L. O'Callaghan, and R. Panigrahy. Better streaming algorithms for clustering problems. *Proceedings of the Symposium on Theory of Computing*, pages 30–39, 2003.

[FIS04] G. Frahling, P. Indyk, and C. Sohler. Estimating the weight of euclidean minimum spanning trees in data streams. *Manuscript*, 2004.

[FKZ02] Joan Feigenbaum, Sampath Kannan, and Jian Zhang. Computing diameter in the streaming and sliding-window models. *Yale University Technical Report YALEU/DCS/TR-1245*, 2002.

[FM85] P. Flajolet and G. Martin. Probabilistic counting algorithms for data base applications. *Journal of Computer and System Sciences*, 31:182–209, 1985.

[GMMO01] S. Guha, N. Mishra, R. Motwani, and L. O'Callaghan. Clustering data streams. *Proceedings of the Symposium on Theory of Computing*, 2001.

[HPM04] S. Har-Peled and S. Mazumdar. Coresets for k-means and k-medians and their applications. *Proceedings of the Symposium on Theory of Computing*, 2004.

[HS04] J. Hershberger and S. Suri. Adaptive sampling for geometric problems over data streams. *Proceedings of the ACM Symposium on Principles of Database Systems*, 2004.

[Ind01] P. Indyk. Tutorial: Algorithmic applications of low-distortion geometric embeddings. *Proceedings of the Symposium on Foundations of Computer Science*, 2001.

[Ind03] P. Indyk. Better algorithms for high-dimensional proximity problems via asymmetric embeddings. *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 2003.

[Ind04] P. Indyk. Algorithms for dynamic geometric problems over data streams. *Proceedings of the Symposium on Theory of Computing*, 2004.

[Mey01] Adam Meyerson. Online facility location. *Proceedings of the Symposium on Foundations of Computer Science*, pages 426–431, 2001.

[MP80] J. I. Munro and M. S. Paterson. Selection and sorting with limited storage. *TCS*, 12, 1980.

[MS03] S. Muthukrishnan and M. Strauss. Maintenance of multidimensional histograms. *Proceedings of the FSTTCS*, 2003.

[Mut03] S. Muthukrishnan. Data streams: Algorithms and applications (invited talk at soda'03). *Available at http://athos.rutgers.edu/∼muthu/stream-1-1.ps*, 2003.

[STZ04] S. Suri, C. Toth, and Y. Zhou. Range counting over multidimensional data streams. *Proceedings of the ACM Symposium on Computational Geometry*, 2004.

[TGIK02] N. Thaper, S. Guha, P. Indyk, and N. Koudas. Dynamic multidimensional histograms. *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD)*, 2002.