Data Streams and Data Synopses for Massive Data Sets
(Invited Talk)

Yossi Matias

Tel Aviv University,
HyperRoll Inc., Stanford University
matias@tau.ac.il

Abstract. With the proliferation of data intensive applications, it has become necessary to develop new techniques to handle massive data sets. Traditional algorithmic techniques and data structures are not always suitable to handle the amount of data that is required and the fact that the data often streams by and cannot be accessed again. A field of research established over the past decade is that of handling massive data sets using data synopses, and developing algorithmic techniques for data stream models. We will discuss some of the research work that has been done in the field, and provide a decades' perspective to data synopses and data streams.

1 Summary

In recent years, we have witnessed an explosion in data used in various applications. In general, the growth rate in data is known to exceed the increase rate in the size of RAM, and of the available computation power (a.k.a. Moore's Law). As a result, traditional algorithms and data structures are often no longer adequate to handle the massive data sets required by these applications.

One approach to handle massive data sets is to use *external memory algorithms*, designed to make an effective utilization of I/O. In such algorithms the data structures are often implemented in external storage devices, and the objective is in general to minimize the number of I/Os. For a survey of works on external memory algorithms see [6]. Such algorithms assume that the entire input data is available for further processing. There are, however, many applications where the data is only seen once, as it "streams by". This may be the case in, e.g., financial applications, network monitoring, security, telecommunications data management, web applications, manufacturing, and sensor networks. Even in data warehouse applications, where the data may in general be available for additional querying, there are many situations where data analysis needs to be done as the data is loaded into the data warehouse, since the cost of accessing the data in a fully loaded production system may be significantly larger than just the basic cost of I/O. Additionally, even in the largest data warehouses, consisting of hundreds of terabytes, data is only maintained for a limited time, so access to historical data may often be infeasible.

It had thus become necessary to address situations in which massive data sets are required to be handled as they "stream by", and using only limited memory. Motivated by this need, the research field of data streams and data synopses has

J. Gama et al. (Eds.): ECML 2005, LNAI 3720, pp. 8 – 9, 2005.

emerged and established over the last few years. We will discuss some of the research work that has been done in the field, and provide a decades' perspective to data streams and data synopses. A longer version of this abstract will be available at [4].

The data stream model is quite simple: it is assumed that the input data set is given as a sequence of data items. Each data item is seen only once, and any computation can be done utilizing the data structures maintained in main memory. These memory resident data structures are substantially smaller than the input data. As such, they cannot fully represent the data as is the case for traditional data structures, but can only provide a synopsis of the input data; hence they are denoted as *synopsis data structures*, or *data synopses* [3].

The use of data synopses implies that data analysis that is dependent on the entire streaming data will often be approximated. Furthermore, ad hoc queries that are dependent on the entire input data could only be served by the data synopses, and as a result only approximate answers to queries will be available. A primary objective in the design of data synopses is to have the smallest data synopses that would guarantee small, and if possible bounded, error on the approximated computation.

As we have shown in [1], some essential statistical data analysis, the so-called *frequency moments,* can be approximated using synopses that are as small as polynomial or even logarithmic in the input size. Over the last few years there has been a proliferation of additional works on data streams and data synopses. See, e.g., the surveys [2] and [5]. These works include theoretical results, as well as applications in databases, network traffic analysis, security, sensor networks, and program profiling; synopses include samples, random projections, histograms, wavelets, and XML synopses, among others. There remain a plethora of interesting open problems, both theoretical as well as applied.

References

1. Alon, N., Matias, Y., Szegedy, M.: The space complexity of approximating the frequency moments. J. of Computer and System Sciences 58 (1999), 137-147. STOC'96 Special Issue
2. Babcock, B., Babu, S., Datar, M., Motwani, R. Widom, J.: Models and issues in data stream systems. In Proc. Symposium on Principles of Database Systems (2002), 1-16
3. Gibbons, P.B., Matias, Y.: Synopses data structures for massive data sets. In: External memory algorithms, DIMACS Series Discrete Math. & TCS, AMS, 50 (1999). Also SODA'99
4. Matias, Y.: Data streams and data synopses for massive data sets. http://www.cs.tau.ac.il/~matias/streams/
5. Muthukrishnan, S.: Data streams: Algorithms and applications. http://www.cs.rutgers.edu/~muthu/stream-1-1.ps
6. Vitter, J.S.: External memory algorithms and data structures. ACM Comput Surv. 33(2): 209-271 (2001)