Guest Editor Introduction: Special Section on Online Analysis and Querying of Continuous Data Streams

Rajeev Rastogi

In a number of application domains, data arrives continuously in the form of a stream and needs to be processed in an online fashion. For example, in the network installations of large Telecom and Internet service providers, detailed usage information (e.g., Call Detail Records or CDRs, IP traffic statistics due to SNMP/RMON polling, etc.) from different parts of the network needs to be continuously collected and analyzed for interesting trends. Other applications that generate rapid, continuous, and large volumes of stream data include transactions in retail chains, ATM, and credit card operations in banks, weather measurements, sensor networks, etc. Further, for many mission-critical tasks such as fraud/anomaly detection in Telecom networks, it is important to be able to answer queries in real time and infer interesting patterns online. As a result, recent years have witnessed an increasing interest in designing single-pass algorithms for querying and mining data streams that examine each element in the stream only once.

The large volumes of stream data, real-time response requirements of streaming applications, and architecture of modern computers impose two additional constraints on algorithms for querying streams: 1) The time for processing each stream element must be small, and 2) the amount of memory available to the query processor is limited. Thus, the challenge is to develop algorithms that can summarize data streams in a concise, but reasonably accurate, synopsis that can be stored in the allotted (small) amount of memory and can be used to provide approximate answers to user queries with some guarantees on the approximation error.

Given the plethora of streaming applications and the nontrivial computational challenges they pose, the timing for a special issue on the topic could not have been better. This special issue of Transactions on Knowledge and Data Engineering presents five papers that propose novel synopses structures and fundamental algorithmic techniques for analyzing and querying continuous data streams. Of the five papers, four explore the space/accuracy tradeoff of stream processing algorithms for important problems like clustering and distinct value estimation, and one addresses issues related to the semantics of query operators on (infinite) streams.

The first paper by Guha et al. is illustrative of a general class of streaming algorithms based on the principle of divide-and-conquer. Conceptually, the algorithm proposed in the paper partitions the input stream into chunks and computes a succinct summary for each chunk. Then, in subsequent steps, it repeatedly combines chunk summaries from the previous step to compute new summaries until the final desired summary for the stream is obtained. Guha et al. show how this divide-and-conquer approach can be used to compute $k$ centers for a stream, where each intermediate summary is a set of $O(k)$ centers. The end result is a deterministic constant-factor approximation algorithm for clustering data streams.

In the second paper, Cormode et al. exploit properties of $p$-stable distributions to estimate, with high probability, the number of distinct elements in a stream. Essentially, given a vector of random variables from a $p$-stable distribution, the $L_p$ norm of a stream can be computed by summing the variables, after weighting each variable with the frequency of the corresponding stream element. Thus, choosing random variables from a $p$-stable distribution with a small $p$ yields the number of distinct values in the stream.

Wavelet transforms have been shown to be effective for approximating the frequency distribution of data. In their paper, Gilbert et al. present a randomized “sketch”-based method for estimating, in a streaming environment, the top few Wavelet coefficients with the highest energy. A key contribution of the paper is a special construction (based on second-order Reed-Muller codes) of the random variables used in sketching, so that Wavelet coefficients (and arbitrary range-sum queries) can be obtained very fast. Furthermore, the random sketch synopses considered in the paper can also be used to estimate join sizes, histograms, quantiles, and frequent elements in a stream.

Interesting questions arise, especially for infinitely long data streams, when we consider the semantics of query operators like joins of two or more streams or group-by operators on a single stream. The paper by Tucker et al. refers to the first category of operators as unbounded stateful operators which are those that need to maintain state with no upper bound in its size and, so, may run out of memory. The latter class of operators belong to the category of blocking operators, which are those that need to read the entire input before emitting a single output and might never produce a result (if the stream is infinite). In order to address the above-mentioned problems posed by unbounded stateful and blocking operators, Tucker et al. enhance the streaming

\* R. Rastogi is with Bell Laboratories, Lucent Technologies, 700 Mountain Ave., Room 2B-301, Murray Hill, NJ 07974.  
E-mail: rastogi@research.bell-labs.com.

For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number 118140.
model by allowing each stream to contain special items called punctuations. Informally, a punctuation marks the end of a subset of data and says that no more tuples will follow that match the punctuation—this enables the amount of state to be reduced and tuples to be output early.

The final paper by Ananthakrishna et al. considers the problem of approximately answering correlated-sum aggregate queries on a data stream. The authors show how generalized sample summaries (which are essentially a set of samples from the input data stream) can be used to approximate answers to the correlated-sum queries. Interestingly, the authors are also able to prove that, if the inner query aggregate cannot be computed exactly, then the query result itself cannot be approximated to within a desired epsilon precision.

In summary, the five papers in the special issue are representative of promising approaches for tackling hard data streaming problems. I am confident that the papers will not only help with laying important algorithmic and semantic foundations for the data streaming field, but also stimulate additional thought and contribute to further advances in the area.

Rajeev Rastogi
Guest Editor

Rajeev Rastogi received the BTech degree in computer science from the Indian Institute of Technology, Bombay, in 1988, and the masters and PhD degrees in computer science from the University of Texas, Austin, in 1990 and 1993, respectively. He is the director of the Internet Management Research Department at Bell Laboratories, Lucent Technologies. He joined Bell Laboratories in Murray Hill, New Jersey, in 1993 and became a distinguished member of technical staff (DMTS) in 1998. Dr. Rastogi is active in the field of databases and has served as a program committee member for several conferences in the area. He currently serves on the editorial board of the IEEE Transactions on Knowledge and Data Engineering. His writings have appeared in a number of ACM and IEEE publications and other professional conferences and journals. His research interests include database systems, network management, and knowledge discovery. His most recent research has focused on the areas of network topology discovery, monitoring, configuration and provisioning, XML publishing, approximate query answering, and data streaming.