Offline and Data Stream algorithms for efficient computation of synopsis structures

Sudipto Guha*
Computer and Information Sciences
University of Pennsylvania.
Email: sudipto@cis.upenn.edu

Kyuseok Shim†
Electrical Engineering and Computer Science
Seoul National University.
Email: shim@ee.snu.ac.kr

Abstract

Synopsis and small space representations are important data analysis tools and have long been used OLAP/DSS systems, approximate query answering, query optimization and data mining. These techniques represent the input in terms broader characteristics and improve efficiency of various applications, e.g., learning, classification, event detection, among many others. In a recent past, the synopsis techniques have gained more currency due to the emerging areas like data stream management.

In this tutorial, we propose to revisit algorithms for Wavelet and Histogram synopsis construction. In the recent years, a significant number of papers have appeared which has advanced the state-of-the-art in synopsis construction considerably. In particular, we have seen the development of a large number of efficient algorithms which are also guaranteed to be near optimal. Furthermore, these synopsis construction problems have found deep roots in theory and database systems, and have influenced a wide range of problems. In a different level, a large number of the synopsis construction algorithms use a similar set of techniques. It is extremely valuable to discuss and analyze these techniques, and we expect broader pictures and paradigms to emerge. This would allow us to develop algorithms for newer problems with greater ease. Understanding these recurrent themes and intuition behind the development of these algorithms is one of the main thrusts of the tutorial.

Our goal will be to cover a wide spectrum of these topics and make the researchers in VLDB community aware of the new algorithms, optimum or approximate, offline or streaming. The tutorial will be self contained and develop most of the mathematical and database backgrounds needed.

About the Speakers

Sudipto Guha is an assistant professor in the Computer Information Sciences Department, University of Pennsylvania. He has previously worked at AT&T Labs – Research from 2000 to 2001 as a member of the technical staff after receiving his PhD from Stanford University. His research interests are primarily in design and analysis of algorithms for computation under constrained resources. He has worked in the areas of graph approximation algorithms for NP-hard problems, randomized algorithms and combinatorial optimization, efficient optimization in database query and mining, and data stream algorithms.

Kyuseok Shim is an Associate Professor at School of Electrical Engineering and Computer Science at Seoul National University, Korea. Previously, he was an Assistant Professor at Computer Science Department of KAIST, a member of technical staff and one of the key contributors to the Serendip data mining project at Bell Laboratories, and a research staff at IBM Almaden Research Center. He has been working in the areas of data mining, semi-structured data (XML), stream data, histograms, query processing, query optimization and data warehousing.

*Supported in part by an Alfred P. Sloan Research Fellowship and by an NSF Award CCF-0430376.

†Supported by the Ministry of Information and Communication in Korea through the University Information Technology Research Center (ITRC) Support Program.

Proceedings of the 31st VLDB Conference, Trondheim, Norway, 2005