

Efficient Approximate Query Processing in Peer-to-Peer Networks

Speaker: Dr. Gautam Das, CSE@UTA

Date: Friday, Sept 29, 2006

Time: 10 am - 12 noon

Venue: WH 413 (CReWMaN Lab)

Abstract

Peer-to-peer (P2P) databases are becoming prevalent on the Internet for distribution and sharing of documents, applications, and other digital media. The problem of answering large scale, ad-hoc analysis queries - e.g., aggregation queries - on these databases poses unique challenges. Exact solutions can be time consuming and difficult to implement given the distributed and dynamic nature of peer-to-peer databases. In this talk, we will present novel sampling-based techniques for approximate answering of ad-hoc aggregation queries in such databases. Computing a high-quality random sample of the database efficiently in the P2P environment is complicated due to several factors - the data is distributed (usually in uneven quantities) across many peers, within each peer the data is often highly correlated, and moreover, even collecting a random sample of the peers is difficult to accomplish. To counter these problems, we have developed an adaptive two-phase sampling approach, based on random walks of the P2P graph as well as block-level sampling techniques. We present extensive experimental evaluations to demonstrate the feasibility of our proposed solution. We will also discuss how to extend this approach to wireless sensor networks.

Biography

Dr. Gautam Das has been with the CSE department of UT-Arlington since Fall 2004, where he is an Associate Professor and Head of the Database Exploration Laboratory (DBXLAB). Prior to joining UT-Arlington, Dr. Das was a Researcher in the Data Management, Exploration and Mining (DMX) Group at Microsoft Research. He has also held positions at Compaq Corporation and the University of Memphis. He graduated with a B.Tech. in computer science from IIT Kanpur, India, and with a Ph.D. in computer science from the University of Wisconsin, Madison.

Dr. Das's research interests span data mining, information retrieval, databases, approximate query processing, distributed, P2P and heterogeneous data management, applied graph and network algorithms, computational biology and computational geometry. While at Microsoft Research, he was extensively involved in the design and implementation of association mining algorithms, sampling based approximate query processing, and in the integration of

database and information retrieval techniques. Prior to Microsoft, he has worked in clustering of categorical data and time series similarity, and also in the classical algorithms areas of shortest paths, spanning networks and geometric visibility. His research has resulted in over 60 papers, many of which have appeared in premier data mining, database and algorithms conferences such as SIGMOD, VLDB, ICDE, KDD, STOC, SODA, and SoCG, as well as in several leading journals and invited book chapters. Dr. Das has served as Best Paper Awards Chair of KDD 2006, Program Co-Chair of CIT 2004 and SIGMOD-DMKD 2004, as well as in program committees of premier conferences such as PODS, SIGMOD, ICDE, KDD, ICDM, SDM, ICML and others