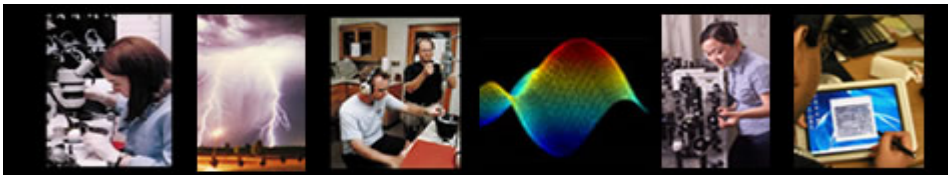
Discovery Park

Cyber Center



LECTURE

March 12, 2010

12:00 Noon

Lawson 1142

Liangcai (Steven) Shu

Efficient Blocking for Entity Resolution

In many telecom and web applications, there is a need to identify whether data objects in the same source or different sources represent the same entity in the real-world. This problem arises for subscribers in multiple services, customers in supply chain management, and users in social networks. It becomes even more difficult to solve when data is integrated from multiple data sources as there usually lacks unique identifier in the system to represent a real-world entity. Entity resolution is to identify and discover objects in the data sets that refer to the same entity in the real-world.

We investigate the entity resolution problem for large data sets where efficient and scalable solutions are needed. We propose an unsupervised blocking method, which is used to divide a data set into blocks such that candidate objects representing the same entity appear in the same block. Our experimental results with real-world data show that our approach is promising.

Bio:

Mr. Liangcai Shu is a Ph.D. candidate in the Department of Computer Science at the State University of New York at Binghamton. His research interests include information retrieval, information extraction, data mining and machine learning. And his PhD dissertation research focuses on the entity resolution problem. He has published research papers in ICDE and WISE, and received the Best Runner-up Student Paper Award in APWeb-WAIM 2009. During the summer of 2009, he did research at Bell Labs as a student intern. Afterwards he continues collaborating with researchers at the labs towards applying for a US patent for their research.

PURDUE UNIVERSITY