IS Department Seminar

Working with Web Data: Semantics and Scalability

Cong Yu

www.eecs.umich.edu/~congy
Research Scientist
Google Research NYC

DATE: Wednesday, October 19, 2011
COFFEE: 2:15-2:30 PM
TIME: 2:30-4:00 PM
LOCATION: GfTC Building, Room 1403

Abstract
The Web has brought unprecedented data challenges in terms of both semantics and scalability. For semantics, the Social Web has produced vast amounts of collaborative ratings and reviews data that present many interesting challenges on how to interpret them to benefit the end users. For scalability, the sheer scale of the Web log data being accumulated at many companies is making tasks such as aggregation analysis much more challenging to perform. In this talk, I outline the major challenges and describe our works in addressing some of them. Specifically, to address the rating data interpretation challenge, we propose efficient algorithms within a novel data cube based mining framework, and show that they can help ordinary users understand ratings data better than simple mechanisms such as average. To address the challenge of aggregation with holistic measures, we propose the MR-Cube system, which is based on the Map Reduce parallel computing model. We show experimentally that the MR-Cube system has orders of magnitudes better scalability than existing techniques. In the final part of the talk, I will briefly describe additional challenges we are currently addressing.

BIO
Cong Yu is a Research Scientist at Google Research in NYC. At Google, Cong works on the WebTables project and is primarily interested in scalable structured data extraction and processing. He also works with various external collaborators to explore research topics in social content mining and database usability. Before joining Google, Cong was a Research Scientist at Yahoo! Research NYC for three years. He obtained his Ph.D. in Computer Science and Engineering from the University of Michigan, Ann Arbor. His dissertation, "Managing Complex Databases in a Schema Management Framework," was an Honorable Mention for the ACM SIGMOD Dissertation Award in 2008. Outside of his work, Cong is a devoted Michigan football fan and is happy that the program is finally on the right track after three years.