IS Department Seminar

Scraping and Processing Financial Documents and Data from the Web

Dr. Quanzhi Li
Engineering Manager at FactSet

DATE: Wednesday November 14, 2012
COFFEE: 2:15-2:30
TIME: 2:30-4:00
LOCATION: GITC Building, Room 1403

ABSTRACT
The number of companies posting financial information, such as earning report and fund holdings, on the Internet has increased exponentially in recent years. Most SEC websites and government regulator sites also publish various financial data, such as bond issues and shareholder changes. Investment professionals rely on these data to make quick and correct decision. It is a challenging task to acquire and process these data with low latency and high quality. In this presentation I will talk about some projects I have worked on and the challenges. These projects involve crawling/scraping the web to obtain various financial data, and processing them to provide information to investment professionals, such hedge fund manager. Related technologies used are web scraping, NLP, IR, information extraction, machine learning and semantic search.

BIO
Dr. Quanzhi Li is an Engineering Manager at FactSet. Before joining FactSet, he worked at FT Search, Financial Times Group as a Senior Software Engineer and Director of Semantic Technology, and at Avaya as a Research Scientist. His industry experience and research interests are in areas of web crawling/scraping, NLP, IR, data and text mining, information extraction, document management, machine learning and semantic technology. He has published over 20 papers in refereed journals and conferences. He obtained his PhD in Information Systems from NJIT, and holds a MS in Computer Science from University of Minnesota, Twin Cities and a BS in Electrical Engineering from Tsinghua University, China.