CS 595-06 Data Provenance 2012

Mon + Wed 3:15 - 4:30 PM, Room: Stuart 106

Instructor: Boris Glavic, Stuart Building 226 C, Phone: 312 567 5205, Email: bglavic@iit.edu

Office Hours: Thursday, 1:00 pm - 2:00 pm

Course Webpage: Will be linked on www.cs.iit.edu/~glavic soon!

Course Description:

With the ever increasing amount of digital information comes an increasing need to understand ”where” an piece of data (data item) is coming from, ”why” it is in the result of a data transformation, and ”how” it was produced by the transformation. For example, biologists use complex digital workflow and simulations to gain new insights from measurement and derived data. The result data of a complex workflow is meaningless without information of how the data was produced from which input data. This type of information, i.e., information about the creation process and origin of data, is called data provenance.

Systems that automatically track provenance information for data produced by e.g., workflows or SQL queries are becoming more and more important. Data provenance is an emerging technology which is used to, e.g., trace errors in transformed data back to its origin or gain additional insights about the data.

We will study several models of provenance developed for domains such as databases and workflow systems. This course covers approaches for automatically tracking provenance, and study query languages and storage mechanism for provenance information. Furthermore, we will discuss real systems that generate provenance data. This course gives the students the opportunity to learn about a hot topic in database research and work with novel research prototype provenance systems.

The course will consist of lectures, reviews of research papers, project presentations by the students, and possibly invited speakers.

Course Material:

No text book is required. Required reading will consist of research publications that are available online.

Prerequisites:

Some background knowledge in databases is required to understand the material. For example, attending one of the courses CS 425, CS 520, or CS 525 should be sufficient. The course is mainly for graduate students. Undergraduates should talk to me first.

Course Details:

The following topics will be covered in the course:

- Introduction to Data Provenance
  - What is Data Provenance?
  - Why do we need it?
  - Understanding different types of data provenance
- Database Provenance
  - Provenance Models and Systems
    - Why-provenance
    - Where-provenance and the DBNotes system
    - Lineage and the WHIPS prototype
    - Witness-list semantics and Perm
    - Provenance semirings and Orchestra
    - Causality and Responsibility models
  - Storage mechanisms
  - Query languages
- Extensions of the Provenance concept
  - Provenance for missing answers
  - Provenance for past queries
  - Provenance for updates
- Beyond Database Provenance
  - Scientific workflows
  - Provenance in the operating system context
  - Connection with Dataflow analysis in programming languages

Project:

Students will choose a term-long project requiring possibly the implementation/extension of a provenance system, a written report, and two oral presentations (one during and one at the end of the semester). Topics will be related to the material discussed in the course. In addition to the topics presented below, students have the opportunity to suggest their own topics.

- Adding new provenance types to existing systems
  - Add causality based provenance to *Perm*
  - Add Why-provenance to *Perm*
  - Extend the How-provenance implementation of *Perm* to derive new information

- Build a visualization tool for provenance in *Perm*
- . . .
- Your topic

Grading Policies:

- Course Project (Implementation, written report, oral presentation): 60%
- Paper reviews: written review (15%) and oral presentation (25%)
- Participation in the paper discussions: (10%)

Course Objectives:
After attending the course students should be able to:

- Understand the concept of data provenance
- Understand different types of data provenance, their application domains, and relationship to each other
- Understand data provenance models; know their advantages and limitations
- Understand generation and storage mechanisms for data provenance
- Build or extend a system for tracking data provenance
- Read, understand, and summarize a research paper