Computer Science
Seminar

Record Linkage: Concepts and Practice with FRIL

Pawel Jurczyk
Emory University

Abstract: Record Linkage is the task of finding entries that refer to the same entity in two or more data sets. It is useful for joining and cleaning data sets that do not have unique common keys, and is important in many studies in business, public health, sociology and psychology, where reconciling independently collected data sets is often the most important and time-consuming first step.

In this talk, we describe the basic concepts of automated record linkage and demonstrate FRIL, an open-source software collaboratively developed at Emory and the CDC that facilitates fast and accurate linkages. FRIL implements an array of standard record matching algorithms as well as a large number of user tunable parameters for improving efficiency and precision. It provides a friendly, easy to use interface and a set of useful visualization tools for quick user feedbacks. Since its release in 2008, FRIL has been widely adopted, including uses by research groups at the National Center on Addiction and Substance Abuse, the Harvard Business School and the NIH.

This talk should be of interest to faculty and graduate students whose research involve data collection, integration, and cleansing.

Pawel Jurczyk is a Ph.D. Candidate in the Computer Science and Informatics Program at Emory University. His interest includes distributed computing, distributed database query processing, and data/information privacy. He has been an ORISE fellow at the CDC since 2007, and has taught courses in Introduction to Computer Science and Databases.

Friday, September 25, 2009, 3:00 pm
Mathematics and Science Center: W301

Mathematics and Computer Science
Emory University