Application of Similarity-based Clustering to Entity Resolution

ABSTRACT

Entity resolution is the process of reconciling data from different sources that represent a distinct entity -- an individual, for example. The entity invariably is represented differently in each source. These differences, combined with data errors, create instances where groups of similar records brought together by pair-wise matching represent more than one entity.

This presentation discusses how new clustering techniques can be used to automate the validation of whether record group in fact represents a single entity, to partition groups that represents multiple entities, grade the quality of the entity resolution process, and flag problem record groups for segregation and study. The method makes use of positive and negative information (for examples records that match and records that should not be in the same group), and can be used to validate, grade, and partition entities represented by weighted, directed and undirected graphs.

BIOGRAPHY

Thomas Schweiger
Acxiom Corporation

Dr. Schweiger is a researcher and consultant with the Acxiom Corporation and is associate director of the Acxiom Laboratory for Applied Research. His interests are in data mining and recognition. Prior to join Acxiom he was a research engineer at E. I. DuPont de Nemours.

Xiaowei Xu
University of Arkansas at Little Rock

Prof. Xu is Professor of Information Science at UALR. He has a wide spectrum of expertise in data mining, databases, machine learning, information retrieval and high-performance computing. He has published over 30 peer-reviewed research papers, and his data mining algorithms have been included in several standard textbooks.