Public Abstract

First Name:Timothy
Middle Name:Craig
Last Name:Havens
Adviser's First Name:James
Adviser's Last Name:Keller
Co-Adviser's First Name:
Co-Adviser's Last Name:
Graduation Term:SS 2010
Department:Computer Engineering
Degree:PhD
Title:Clustering in Relational Data and Ontologies

This dissertation studies the problem of clustering objects represented by *relational data*. This is a pertinent problem as many real-world data sets can only be represented by relational data for which object-based clustering algorithms are not designed. Relational data are encountered in many fields including biology, management, industrial engineering, and social sciences. Unlike *numerical object data*, which are represented by a set of feature values (e.g. height, weight, shoe size) of an object, relational object data are the numerical values of *(dis)similarity* between objects. For this reason, conventional cluster analysis methods such as K-means and fuzzy c-means cannot be used directly with relational data.

I focus on three main problems of cluster analysis of relational data: (i) *tendency prior to clustering*---how many clusters are there?; (ii) *partitioning of objects*---which objects belong to which cluster?; and (iii) *validity of the resultant clusters*---are the partitions ``good''? Analyses are included in this dissertation that prove that the *Visual Assessment of cluster Tendency* (VAT) algorithm has a direct relation to *single-linkage* hierarchical clustering and *Dunn's cluster validity index*. These analyses are important to the development of two novel clustering algorithms, **CLODD**-*CLustering in Ordered Dissimilarity Data* and **ReSL**-*Rectangular Single-Linkage* clustering.

Last, this dissertation addresses clustering in ontologies, a specific type of data; examples include the Gene Ontology, the MeSH ontology, patient medical records, and web documents. I apply an extension to the Self-Organizing Map (SOM) to produce a new algorithm, the **OSOM**-*Ontological Self-Organizing Map*. OSOM provides visualization and linguistic summarization of ontology-based data. A binary-valued vector-based network prototype is used to represent ontological objects (e.g. genes and gene products).