ClusterGram – A visualization tool to aid data clustering and entity resolution

Ren Bauer
University of North Carolina at Chapel Hill

Abstract
When performing entity resolution and clustering large datasets, it can be difficult to get a feel for where to begin. Even after the creation of a distance matrix, determining the optimal parameters for generating optimal clusters can be a daunting challenge. ClusterGram is a light java application that parses distance matrices in an attempt to provide a concise visualization from which useful information can be obtained. In specific, the application strives to provide approximate distance measure cutoffs which can be used as guidelines when clustering data and performing entity resolution.

Overview
ClusterGram employs a visualization technique akin to a circular dendrogram to display distance matrices on a single plane:

Using this novel technique, distance is represented both horizontally and vertically. Thus, large sets of data can be represented in a visually pleasing view that allows users to extract valuable information.

Evaluation
By using ClusterGram to view datasets with known clusters, it’s ability to provide effective visualizations of those clusters becomes clear:

Note: In this example, the two blue clusters are distinct

Features
Zoom and Pan
Rotate
Guide Entity Resolution
Guide Clustering
3 Cluster Distance Metrics