Title: A Self-Tuning Diffusion Map Framework for Use in Document Clustering

Abstract:

Document clustering has been a rich research area, resulting in algorithms for grouping a fixed or streaming corpus when topic labels are unknown or pre-defined. Regardless of approach, most methods suffer from the need to analyze a very high-dimensional space of words in the corpus lexicon. This dimensionality is often reduced prior to analysis via some statistical threshold or common sense heuristic (e.g. removing words like "the"). It might be beneficial to remove this somewhat subjective decision. Diffusion maps are a powerful tool for identifying complicated structure and reducing dimensionality in a wide variety of applications. Representing the connectivity of a data set, diffusion maps project observations into a space in which standard methods can more easily model the structure. We explore the use of a flexible self-tuning diffusion map framework that incorporates local tuning parameters to capture group structure of varying density, if present, in a corpus of documents. Our work thus far has also shown a decrease in importance of the clustering method choice once in the reduced projected space. Although the primary focus of this talk is the recovery of cluster structure, we also present classification and regression frameworks.