Semantic Based Clustering of Web Documents

蔣以仁

Tsau Young Lin;I-Jen Chiang

摘要

Abstract

A new methodology that structures the semantics of a collection of documents into the geometry of a simplicial complex is developed: a primitive concept is represented by a top dimension simplex, and a connected component represents a concept. Based on these structures, documents can be clustered into some meaningful classes. Experiments with three different data sets from web pages and medical literature have shown that the proposed unsupervised clustering approach performs significantly better than traditional clustering algorithms, such as k-means, AutoClass and hierarchical clustering (HAC). This abstract geometric model seems have captured the intrinsic semantics of the documents.