World Academy of Science, Engineering and Technology
International Journal of Mathematical and Computational Sciences
Vol:14, No:12, 2020

A Clustering Algorithm for Massive Texts

Authors : Ming Liu, Chong Wu, Bingquan Liu, Lei Chen

Abstract : Internet users have to face the massive amount of textual data every day. Organizing texts into categories can help users dig the useful information from large-scale text collection. Clustering, in fact, is one of the most promising tools for categorizing texts due to its unsupervised characteristic. Unfortunately, most of traditional clustering algorithms lose their high qualities on large-scale text collection. This situation mainly attributes to the high- dimensional vectors generated from texts. To effectively and efficiently cluster large-scale text collection, this paper proposes a vector reconstruction based clustering algorithm. Only the features that can represent the cluster are preserved in cluster's representative vector. This algorithm alternately repeats two sub-processes until it converges. One process is partial tuning sub-process, where feature's weight is fine-tuned by iterative process. To accelerate clustering velocity, an intersection based similarity measurement and its corresponding neuron adjustment function are proposed and implemented in this sub-process. The other process is overall tuning sub-process, where the features are reallocated among different clusters. In this sub-process, the features useless to represent the cluster are removed from cluster's representative vector. Experimental results on the three text collections (including two small-scale and one large-scale text collections) demonstrate that our algorithm obtains high quality on both small-scale and large-scale text collections.

Keywords : vector reconstruction, large-scale text clustering, partial tuning sub-process, overall tuning sub-process
Conference Title : ICSRD 2020 : International Conference on Scientific Research and Development
Conference Location : Chicago, United States
Conference Dates : December 12-13, 2020