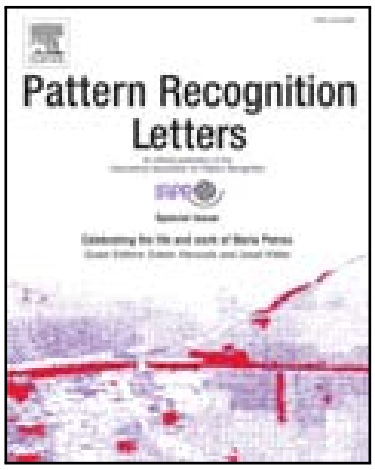

A Fast Partitional Clustering Algorithm based on Nearest Neighbours Heuristics

Debasis Ganguly

PII: S0167-8655(18)30314-3
DOI: 10.1016/j.patrec.2018.07.017
Reference: PATREC 7245

To appear in: *Pattern Recognition Letters*

Received date: 7 March 2018
Revised date: 1 June 2018
Accepted date: 14 July 2018

Please cite this article as: Debasis Ganguly, A Fast Partitional Clustering Algorithm based on Nearest Neighbours Heuristics, *Pattern Recognition Letters* (2018), doi: 10.1016/j.patrec.2018.07.017

Highlights

- Investigate K-means clustering for large collections of sparse vectors of high dimensionality.
- Proposed utilizing the inverted list data-structure to improve run-time of K-means.
- Heuristics proposed for initial centroid selection and centroid updates.
- Proposed approach outperforms the run-time of K-means by up to 35x on a collection of 14M tweets.