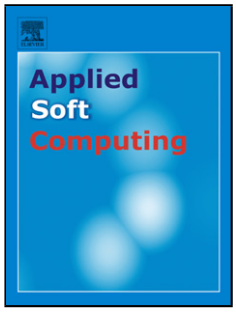

Title: Parallel Swarm Intelligence Strategies for Large-scale Clustering based on MapReduce with Application to Epigenetics of Aging

Author: Zakaria Benmounah Souham Meshoul Mohamed Batouche Pietro Lio'

PII: S1568-4946(18)30203-5
DOI: https://doi.org/doi:10.1016/j.asoc.2018.04.012
Reference: ASOC 4816

To appear in: *Applied Soft Computing*

Received date: 29-4-2016
Revised date: 26-3-2018
Accepted date: 10-4-2018

Please cite this article as: Zakaria Benmounah, Souham Meshoul, Mohamed Batouche, Pietro Lio', Parallel Swarm Intelligence Strategies for Large-scale Clustering based on MapReduce with Application to Epigenetics of Aging, *<![CDATA[Applied Soft Computing Journal]]>* (2018), https://doi.org/10.1016/j.asoc.2018.04.012

Parallel Swarm Intelligence Strategies for Large-scale Clustering based on MapReduce with Application to Epigenetics of Aging

Zakaria Benmounah[1], Souham Meshoul[1], Mohamed Batouche[1], Pietro Lio'[2]

Abstract

Clustering is an important technique for data analysis and knowledge discovery. In the context of big data, it becomes a challenging issue due to the huge amount of data recently collected making conventional clustering algorithms inappropriate. The use of swarm intelligence algorithms has shown promising results when applied to data clustering of moderate size due to their decentralized and self-organized behavior. However, these algorithms exhibit limited capabilities when large data sets are involved. In this paper, we developed a decentralized distributed big data clustering solution using three swarm intelligence algorithms according to MapReduce framework. The developed framework allows cooperation between the three algorithms namely Particle Swarm Optimization, Ant Colony Optimization and Artificial Bees Colony to achieve largely scalable data partitioning through a migration strategy. This latter reaps advantage of the combined exploration and exploitation capabilities of these algorithms to foster diversity. The framework is tested using Amazon Elastic MapReduce service(EMR) deploying up to 192 computer nodes and 30 gigabytes of data. Parallel metrics such as speed-up, size-up and scale-up are used to measure the elasticity and scalability of the framework. Our results are compared with their counterparts big data clustering results and show a significant improvement in terms of time and convergence to good quality solution. The developed model

[☆]Fully documented templates are available in the elsarticle package on CTAN.
[1]Computer Science Department, Constantine 2 University, Constantine, Algeria.
[2]Computer Science Department, Cambridge University, Cambridge, UK.