PENNSTATE

Technology Available for Licensing
Office of Technology Management-The Pennsylvania State Univ.
113 Technology Center, University Park, PA 16802
814.865.6277 phone; 814.865.3591 fax

Non-Confidential Description - PSU No. 3982
"Parallel D2-Clustering: Large-Scale Clustering of Discrete Distributions"

Keywords:

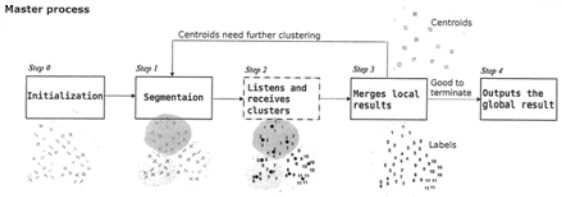
D2 clustering, parallel algorithm, image annotation, big data

Links:

Inventor Website
Published Patent Application

Inventors:

James Wang, Jia Li, Yu Zhang



Parallel Processing

Background

Clustering is a fundamental unsupervised learning methodology for data mining and machine learning. The D2 clustering algorithm applies to image annotation and is constructed based on the Mallows distance, which provides a metric for image retrieval and annotation. Scalability has emerged as a problem with D2 clustering because the amount of unknown variables grows with the number of objects in a cluster. As a result, it takes several minutes to learn each category by performing D2 clustering on 80 images, and would take more than a day to complete the modeling of thousands of images.

Invention Description

This algorithm uses a divide-and-conquer strategy in a novel parallel algorithm to reduce the computational complexity of D2 clustering. The goal is to parallelize the centroid update in D2 clustering by: dividing the data into segments based on their adjacency, computing some local centroids for each segment in parallel, and combining the local centroids to a global centroid. This parallel algorithm achieves significant speed up with minor accuracy loss. The computational intensiveness of D2 clustering limits its usage to only relatively small scale problems. With emerging demands to extend the algorithm to large-scale datasets (online image datasets, video resources, and biological databases) this invention exploits parallel processing in a cluster computing environment in order to overcome the inadequate scalability of D2 clustering.

Advantages / Applications

- Speeds up computational time with minor accuracy loss
- Can be applied to large-scale datasets
- The parallel algorithm reduces the computational complexity of D2 clustering

Contact: **Bradley A. Swope**
Sr. Technology Licensing Officer
The Pennsylvania State University

Phone: (814) 863-5987
Fax: (814) 865-3591
E-mail: bradswope@psu.edu

July-15