CATS: A Big Network Clustering Algorithm based on Triangle Structures

Mincheol Shin[1, *], Jeongwoo Kim[1, *], Jungrim Kim [1]

Dongmin Seo [2], Chihyun Park [2], Seok Jong Yu [2], Sanghyun Park [1, †]

{smanioso, jwkim2013, kimgogo02}@yonsei.ac.kr, {dmseo, chihyun.park, codegen}@kisti.re.kr,
sanghyun@yonsei.ac.kr

ABSTRACT

A huge amount of data, known as "big data," has been generated from various areas. A network is a popular data structure for presenting and analyzing big data. However, the conventional network analysis algorithms cannot cover the size of big data. To address this limitation, we propose in this paper a network clustering algorithm for a big data network using a parallel distributed computation model. To consider parallel computation concepts, we change the paradigm of the conventional clustering algorithm using triangle structures. We demonstrate that the proposed algorithm can cover a big data network that cannot be otherwise implemented using a conventional algorithm. Experimental results show that the proposed algorithm is faster than the conventional algorithm.

CCS Concepts

• Theory of computation → Design and analysis of algorithms →Distributed algorithms

Keywords

Network, Clustering, Parallel Distributed Computation, Big Data

1. INTRODUCTION

Recently, the term "big data" has been widely used to describe a huge amount of data. These data have been generated in several areas, such as biology, marketing, social media, and wireless sensor data analysis. Owing to the size of big data, conventional algorithms are unsuitable to process them. To address this challenge, several studies employed a MapReduce programming model as a big data analysis model. MapReduce is a parallel distributed computation model that employs map and reduce functions in the manner of batch processing. By using the MapReduce model, we can solve some big data problems. However, the MapReduce model is inefficient for an iterative mining algorithm because the model must read and write the result of each iteration, which requires a significant amount of time, for each iterative step. This issue can be solved by using Apache Spark as a parallel distributed model. Apache Spark applies a computing paradigm consisting of transformations of immutable data, which is called a resilient distributed dataset (RDD). By using RDD, Spark resolves limitations in the MapReduce computing paradigm. Spark additionally achieves an efficient fault-tolerance scheme using lineage, which is a sequence of transformations. When a failure occurs, Spark recalculates the failed RDD from ancestral RDDs with a lineage [1]. We can therefore design efficient iterative mining algorithms using Spark.

Meanwhile, a network can be used to describe many types of big data with edges and nodes. Furthermore, a network provides opportunities to analyze data by applying several analytical measures, such as centralities, which include degrees, closeness, betweenness, eigenvectors, clustering, and propagation. Of these measures, clustering is widely used to analyze a network in several fields. In the case of biology, we can infer protein modules by analyzing a protein–protein interaction network. In addition, social groups can be identified by using network clustering in the social network. For this reason, several studies have been conducted to analyze the network data by using network clustering algorithms.

In this paper, we propose the CATS clustering algorithm to analyze a big data network in Spark. The proposed algorithm is based on a clustering algorithm using structure similarity. The goal of this study is to develop a clustering algorithm to process big data in Spark. To this end, we designed a clustering algorithm using a parallel distributed concept. In addition, we changed the algorithm paradigm using a triangle structure.

The main contributions of this work include:

- Developing a clustering algorithm for big data network analysis in Spark.

- Using a triangle structure to calculate structure similarity.

[1] Department of Computer Science, Yonsei University, Seoul, Korea
Tel: +82-2-2123-7757

[2] Korea Institute of Science and Technology Information, Daejeon, Korea
Tel: +82-42-869-1796

[*] These authors equally contributed to this paper.

[†] Corresponding author

SAC 2017, April 03-07, 2017, Marrakech, Morocco

DOI: http://dx.doi.org/10.1145/3019612.3019893