Community Detection using Label Propagation on Social Network Data

Ankit Garg, Vikram S Bisht

Computer & Information Science & Engineering, University of Florida

{agarg, vsbisht}@cise.ufl.edu

Abstract

Working on big data is challenging and interesting. Size of the Social Networking data is growing bigger everyday and analysis on such data is getting more complicated. Twitter which has 500 million users has an impressive growth rate. Social network can be viewed as a map of the individuals, and the ways they are related to each other. We exploit this mapping paradigm to use it in our analysis. MapReduce is a buzzword in Big Data industry and a perfect framework to perform analysis on such data. We used Hadoop and Amazon Elastic Map Reduce Cluster for our analysis on Community Detection using Label Propagation Algorithm. Different notion of analysis can be derived from this work.

Objective

To detect communities using the edge data of user connections in social networks using Map Reduce like framework.

DATA SET

- Social Graphs format
  - USER is FOLLOWER
- Directional or Undirectional
- Kaist Dataset for twitter data
  - Huge data 26.4GB
- Graphs generated by LFR benchmark
  - Open source code for generation of Graph DB
  - Testing accuracy using artificially generated communities

Label Propagation Algorithm

1. Initialize labels at all nodes in the network. For a given node x, C(r(x)) = x.
2. Set t = 1.
3. For each x ∈ X chosen in that specific order, let Ct(x) = f(Ct−1(x)) for mappers and reducers can be controlled
   - CLI and Web Interface Available
- BOTO
  - Python integrated interface to Amazon Web Services
- Command line or python script to launch map reduce jobs on EMR

Correctness on LFR benchmark

- LFR benchmark (A. Lancichinetti, S. Fortunato, and F. Radicchi) used to create artificial graphs
- Generated graph data using graph bench
  https://github.com/itsukubato/tree/master/graphdb-bench
- Community detection algorithms: a comparative analysis (Andrea Lancichinetti and Santo Fortunato)
- Generated directional graph using benchmark tool having 20 groups, 1000 nodes and around 2000 edges
- Results 22 groups after 11 iterations

Results

- LFR Benchmark
  - Fig: % Reduction in Communities with Label Propagation in each iteration
  - Table: Results of LFR Benchmark

- Twitter Data
  - % Reduction in Comm. Detection for every iteration
  - Fig: % Reduction in Communities with Label Propagation in each iteration

- Table: Twitter Data

References


Acknowledgements

- We thank Dr. Daisy Zhe Wang (daisyw@cise.ufl.edu), for her guidance and support during the entire course of the project.