Database Design and Implementation

Spring 2014

Project Proposal

Submitted by:

Vijay Kumar Pasikanti, Ashish Jain

February 26, 2014

Problem statement:

In this project, we are presenting a comprehensive study to compare graph databases performance on a graph clustering algorithm. The data will be generated through synthetic graph data generators. We have specified this generator in the last section of proposal.

Goal and motivation:

Graph data management has been receiving a lot of attention with the arrival of various social networks like facebook, twitter, google+ etc. All these networks have their data stored in some or the other form of graph database, which can scale up to billions of nodes. Graph databases process relationship information between nodes and their efficiency running queries making Relational databases obsolete. Social networks are flooded with online communities, which are formed based on some similarity between the users.

However, there is no comprehensive study to measure performance of community detection algorithms in the context of various graph databases.

Therefore, we have two motivations for doing this project:

1. Clustering a synthetic social network graph on graph databases to detect network communities. For accomplishing this task, we will be implementing standard agglomerative clustering with various cluster quality measures to form clusters.

2. We will do the performance evaluation of graph databases in terms of time of execution and memory usage for the clustering algorithm. It will provide a benchmark for future users and an opportunity to select the appropriate graph database based on how its performance metrics.

3. We also would like to study the scalability of graph databases as data grows in orders of magnitude using clustering as the task.

In one of the benchmarking task [8], author has compared performance of Page Rank and Shortest Paths on graph database and relational database. Therefore our proposal is novel in terms of task compared.

Ambitious goal:

If time permits, we plan to implement and evaluate our algorithm on PEGASUS: A peta-scale graph mining system.

A survery of related work:

Hierarchical clustering algorithm has been studied([1][2][3]) for over decades now. Also there have been many of its successful implementations for various applications([4][5]). In addition to that, many researchers have evaluated different types of hierarchical algorithms (list of algo survey papers) but most of the results were presented on smaller datasets. As the data sizes grow for network graphs these algorithms are implemented on graph databases and there is a very little literature ([6]) on the implementation and evaluation of these clustering algorithms on large datasets using graph database systems. The goal of the project is to study the performance of graph database systems running clustering algorithms on large datasets.

Milestones:

1. Dataset generation - 03/03
2. Graph database understanding and hands on: 03/07
3. System setup: 03/12
4. Clustering Algorithm implementation and testing on 1st database system: 03/24
5. Testing on the 1st database: 03/30
6. Clustering Algorithm on 2nd database system: 4/10
7. Complete evaluation: 4/20

System:

- **Graph Databases**: Neo4j, GraphLab
- **Datasets**: NetworkX[9], a Python language software package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks.

References:

1. William H. E. Day, Herbert Edelsbrunner "Efficient algorithms for agglomerative hierarchical clustering methods"

2. Chris Fraley "Algorithms for model-based Gaussian hierarchical clustering"

3. CF Olson "Parallel algorithms for hierarchical clustering"

4. M Steinbach, G Karypis, V Kumar "A comparison of document clustering techniques"

5. Y Zhao, G Karypis "Evaluation of hierarchical clustering algorithms for document datasets"

6. S Guha, R Rastogi, K Shim "CURE: an efficient clustering algorithm for large databases"

7. J. McAuley and J. Leskovec. "Learning to Discover Social Circles in Ego Networks"

8. http://istc-bigdata.org/index.php/benchmarking-graph-databases/

9. http://networkx.github.io