Database Design and Implementation

Spring 2014

Project Proposal

Submitted by:

Vijay Kumar Pasikanti, Ashish Jain

{vijaykp,ashishjain}@cs.umass.edu

February 27, 2014

Problem statement:

In this project, we are presenting a comprehensive study of performance of graph databases on a graph clustering algorithm. The data will be generated through a synthetic graph data generator[9]. We have specified this generator in the last section of proposal.

Goals and motivation:

Graph data management systems have 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 with billions of nodes. Graph databases processing relation information between nodes effectively and their efficiency on running graph queries have made relational databases not a preferred choice. Social networks are flooded with online communities, which are formed based on some form of connectivity between the users.

However, there have been no comprehensive studies in the past 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 performance evaluation of graph databases in terms of their time of execution and run-time memory usage while running the clustering algorithms. It will provide benchmark numbers and an opportunity to select an appropriate graph database.

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

In one of the benchmarking study[8], author has compared performance of graph database and relational databases on page rank and shortest paths tasks. Therefore our proposal is novel in terms of task.

Ambitious goal:

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

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 algorithm 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. Some work on performance comparison on the relational and graph databases has been done by [8] but they evaluated different tasks i.e. page ranking and shortest path distance. They showed that relational databases perform better in comparison to graph database for the above tasks. However when dealing with social network data, SQL is certainly not a preferred choice. Therefore a comprehensive performance comparison of some popular graph databases is required.

Milestones:

1. Dataset generation: 03/05
2. Graph databases understanding and hands on: 03/13
3. System setup: 03/16
4. Clustering algorithm implementation and testing on 1st database system: 03/26
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