Clustering for Graph Partitioning

Andrea Smith

May 3, 2018

About the Author

The author has earned her Bachelor's of Science degree in Computer Science from Missouri University of Science and Technology. She is working towards her doctorate degree in Computer Science, with a focus on graph data mining, from the same university, anticipating completing 2022.

Contents

1 Executive Summary

This project aims to explore the potential of clustering algorithms for aiding in the partitioning of large graphs for frequent subgraph mining. Partitioning is important as many graphs of interest cannot be held in the memory of a single machine, and so must be spread across many machines in order to be processed. Partitioning risks losing data that spans across multiple partitions, requiring an intelligent partitioning process in order to minimize.

This project explores k-means clustering, distance-k clustering, and multilevel kernel clustering algorithms, as they apply to partitioning large transaction graphs in data mining. The clusters produced by the algorithms are evaluated for quality based on cluster density and the maximal shortest path within the cluster. The algorithms were tested on a toy data set built from publicly available Arabian horse pedigree papers, such that each edge represents a child-parent relationship. The algorithms were written for undirected, weighted graphs, but could be modified for directed graphs. Furthermore, research needs to be done to create a parallelized or distributed version of these algorithms so that they can process graphs that do not fit within the memory of a single machine.

2 Introduction

Stuff to say [1, 2].

3 Project Specifications

This project focuses on comparing the quality of the clusters produced by each algorithm. Quality is a combined score of the cluster density and the maximum shortest path within the cluster. By comparing the clustering ability of the algorithms on a small graph, inferences about the quality of partitions made on a larger transaction graph can be made. Note that further research should be done into parallelized and distributed versions of these algorithms, but that is beyond the scope of this project.

4 Detailed Design

5 Experimental Results

6 Appendix A

References

[1] I. Dhillon, Y. Guan, and B. Kulis, "A Fast Kernel-based Multilevel Algorithm for Graph Clustering," *KDD*. Chicago, Ill.: 2005.

[2] J. Edachery, A. Sen, and F. Brandenburg, "Graph Clustering Using Distance-k Cliques," Arizona State University, 1999.

[3] K. Nazeer and M. Sebastian, "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm," *Proceedings of the World Congress on Engineering*. London, U.K.: 2009.