

CSCI4390/6390 – Data Mining

Fall 2007, Group Project

1 Project Description

The project consists of implementing a given algorithm taken from the current literature. After reading the paper, you will implement the algorithm, and you will redo the experiments as well. The whole idea of the project is to be able to replicate the results reported in the paper. You will submit a final report that documents your implementation and experiments.

Here is the list of suggested projects/papers. The papers describing the algorithms can be downloaded from: <http://www.cs.rpi.edu/~zaki/dmcourse/projects/>

1. MARGIN: Maximal Frequent Subgraph Mining (Margin-short.pdf, Margin-long.pdf)
2. SPIN: Mining Maximal Frequent Subgraphs from Graph Databases (Spin-long.pdf)
3. k-means projective clustering (ProjectiveClustering.pdf)
4. SCAN: A Structural Clustering Algorithm for Networks (Scan.pdf)
5. GraphScope: Parameter-free Mining of Large Time-evolving Graphs (GraphScope.pdf)
6. Fast and Practical Indexing and Querying of Very Large Graphs (Gripp.pdf)

Please select one of the topics from above and email me a prioritized list of preferences before class on Thursday, 18th October.

The above topics are primarily on graphs and projected clustering. You may also choose other topics from the last three years of the following conferences of interest:

- SIGKDD: <http://www.informatik.uni-trier.de/~ley/db/conf/kdd/index.html>
- SDM: <http://www.informatik.uni-trier.de/~ley/db/conf/sdm/index.html>
- SIGMOD: <http://www.informatik.uni-trier.de/~ley/db/conf/sigmod/index.html>
- VLDB: <http://www.informatik.uni-trier.de/~ley/db/conf/vldb/index.html>
- ICDM: <http://www.informatik.uni-trier.de/~ley/db/conf/icdm/index.html>
- ICDE: <http://www.informatik.uni-trier.de/~ley/db/conf/icde/index.html>

If you do want to choose another topic, please discuss with me during class on Thursday, 18th October.

2 Project Guidelines

1. Due date is **Monday 19th November**. The project report must be submitted in hard copy during class, and the tar/zip file for the code must be submitted before class.
2. Everyone is expected to use C++ for the implementation, preferably using STL, and the code should compile with a g++ compiler.
3. For the project report, a description in the following format should be submitted:
 - (a) Problem Introduction
 - (b) Algorithm Design (high level idea, pseudo-code)
 - (c) Implementation Issues (data structures used, etc.)
 - (d) Experiments on a number of datasets with different parameters.
 - (e) Future work, challenges overcome and conclusions.

I expect a report of around 10 pages.

4. You must submit your entire source code directory (zipped and tarred). It must have:
 - (a) All *.h and *.cc/*.cpp files
 - (b) A README file on how to run the programs
 - (c) A Makefile to make the sources
 - (d) A script file for running the experiments you used in the report.
5. Feel free to use the STL library or any other publicly available data structures/algorithms library that will help you get your job done fast. However, you must implement the main algorithms yourself.