Course Project

The course project will consist of three parts: Part 1 (submissions), Part 2 (presentations), and Part 3 (final reports for graduate and online students).

Objectives:

Projects are intended to give students the opportunity to explore ideas or directions in data mining (e.g., frequent itemset mining, clustering, classification, ranking, recommendation, similarity search), to discover interesting pattern and knowledge from available application datasets.

Project submissions will include your data and code, configuration, experiment results including figures and tables, and readme file to state how to run your code. Please first use the same datasets and evaluation measures used in corresponding papers. For undergraduate projects and graduate projects which do not provide available datasets and evaluation measures, use the attached two datasets (**CA-GrQc-1.txt** and **com-dblp.ungraph.txt**) as your data matrices or adjacency matrices. Alternative evaluation measures are Dunn index, Silhouette coefficient, and Davies-Bouldin index for clustering, Macro-F1, Micro-F1, and Hamming Loss for classification, MAE and RMSE for recommendation.

Project presentations (20 minutes for each project) will include problem definition, algorithm introduction, program demonstration, experiment results, and Q/A.

Final reports for graduate students are to be up to 5 pages in length and will include the particular problem to be considered, the algorithm to be implemented, the datasets to be evaluated, where and how to get the data, the evaluation measures to be reported, the main ideas of the algorithm, the main steps of the algorithm, the configuration of experiment environment, the analysis of the experiment results, the strong points of the algorithm, and the potential weak points of the algorithm. Final reports are also expected to show some new ideas about extensions of existing data mining algorithms or to develop new algorithms to solve real world problems.

Requirements:

Your submission should address the following issues:
Implement the source codes by yourselves
Effectiveness test on one small dataset: report two scores of evaluation measures
Efficiency test on one small dataset: report running time
Scalability test on one large dataset (optional): report two scores of evaluation measures, report running time

Requirements:

Your presentation and final report should clarify the following issues:
What is the problem and datasets addressed by the project?
What is the data mining algorithm implemented?
What are the knowledge or pattern discovered from the project?
What are your evaluations and ideas for extensions and improvements?

Evaluation:

Every member in a team gets the same score. No copying or sharing of source codes from paper authors or other people is allowed. The basic features of graduate students will be normalized into the range between 0 and 15 points (for graduate) or 30 points (for undergraduate) on a pro rata basis.

Basic features:
Project submissions (20 points)
Project presentations (10 points)
Final reports (10 points, graduate only)

Bonus features:
Scalability test on one large dataset (5 points)