Internship / master project topic: construction of a rich large-scale data network

Jan Ramon

December 2015

1 Context

In order to validate graph based machine learning and data mining research, many benchmarks exist. Unfortunately, most publicly available benchmarks concern unlabeled undirected graphs, and the few which concern heterogeneous data with a rich set of features are usually difficult to access.

2 Objective

This project aims at generating a publicly available dataset which is

- significant, i.e., is useful to solve real-world challenges,
- easy to access, especially for learning/mining algorithms,
- heterogeneous, i.e., containing data from many different relations
- rich, i.e., objects have a rich set of features describing them

This topic comes in several flavors, depending on the domain of choice, e.g., biology, scientific language or technology. In some of these domains, several data collection efforts already exist, and in that case this project will aim at integration and creatively offering new types of data. There is significant liberty for the student to select a domain of interest as long as the primary objectives can be realized.

3 More information

The project would involve

- choose a domain
- map existing data repositories

- detect opportunities, e.g., easy to fill gaps or currently hard to exploit interdisciplinary relationships.
- Design a dataset and their relationship to existing efforts
- design interfaces to common graph based machine learning / data mining algorithms
- implement the dataset and interface
- construct documentation to facilitate usage
- perform an exploratory analysis