SNA 2011 Fall - Assignment 4

Collaboration Prediction in Heterogeneous Social Networks

Instructor: Shou-De Lin (sdlin@csie.ntu.edu.tw)
TA: Jing-Kai Lou, San-Chuan Hung, and Wei-Shih Lin

Goal

In this assignment, you are requested to predict collaboration connections between authors. You will be given a heterogeneous social network about paper publication, and has to predict the co-authorship in the future.

Dataset

The schema of heterogeneous social networks you will be given is shown as follows:

This is a heterogeneous network. There are three types of node in network: author, paper, and conference. Authors write paper, and papers are published by conferences. The whole data set contains 41 conferences from 2000 to 2009.

We split the network into two parts as revealed data and unseen data by time. All the information from 2000/1/1-2007/12/31 are revealed, and you will have to predict co-author relationships for the two unseen years during 2008/1/-2009/12/31. Similar to previous competition, you can submit your predicts to the online leader-board. This leader-board will reveal how well your outputs are for a randomly select 50% of the unseen data, while eventually you will be ranked based on the performance on all unseen data.

The basic statistical information of the training data is shown as below:

train	**2000-2007**
# author	148718
# papers	113385
# co-author relationship	412509

Definition of co-authorship: If two authors published at least one paper together, then we will regard them as co-authors. If two authors published at least one paper together during 2008~2009, we will regard them as co-author between 2008~2009.

There are two type of co-author relationship:

R1. Two persons who are co-authors before 2008, and co-author again during 2008~2009.

R2. Two persons who are NOT co-authors before 2008, and co-author at least one paper between 2008~2009.

You are told that in this dataset, there are about 20k~30k R1 pairs (or links) and 20K~30K R2 links.

Task 1 (Required) : Predict R1 relationship

Some authors wrote paper together before 2008, and they collaborated again after 2008. Please predict such authors pairs. Upload your results in a text file with file-name <Task1-TEAM-{INDEX}.txt>

Each line should contain two author ids separated by a white space. Make sure do not have multiple identical pairs of author nodes.

Example:
===[filename: Task1-TEAM-3.txt]===
a3052 a12307
a1 a343
a7878 a5566
…

Task 2 (Optional) : Predict R2

Please predict the authors who **didn't collaborate** before 2008, but they do co-author with each other during 2008-2009. Upload your results in a text file with file-name <Task2-TEAM-{INDEX}.txt>.

Each line should contain two author ids separated by a white space. Make sure do not have multiple identical pairs of author nodes.

Note that this task is optional. Your team will get extra bonus score if you do well in this task.
1st place team: +15 pts
2nd place team: +10 pts
3rd place team: +5 pts
Reasonable discussion/analysis about your approach for this problem in the report: +5pts

Example:
===[filename: Task2-TEAM-3.txt]===
a3052 a12307
a1 a343
a7878 a5566

Evaluation

We will evaluate your result by F-measure. The formulas are defined as below:

	actual class (expectation)	
predicted class (observation)	**tp** (true positive) Correct result	**fp** (false positive) Unexpected result
	fn (false negative) Missing result	**tn** (true negative) Correct absence of result

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

F-measure is the harmonic mean of precision and recall:

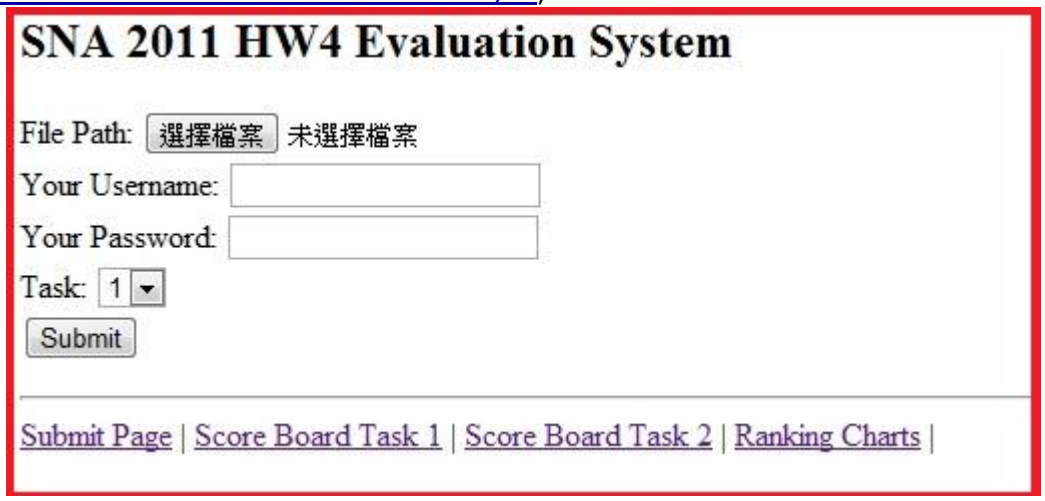
$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}.$$

reference: http://en.wikipedia.org/wiki/F1_scores

Your final grade will be based on the competition ranking (50%) and the report (50%). Your report will be judged by novelty, efforts, and the completeness of works.

Online Leader-board

We have created an online leader-board that allows you to judge how well you do (http://mslab.csie.ntu.edu.tw/SNA2011/EvalSys4).



For each task, each team is allowed to submit your results 15 times per day (starting 11/25 0:01 until 12/14 23:59). Please carefully process your upload. The leader-board will show the ranking of uploaded results, and the last submission before the deadline will be considered as the final submission of your team.

Submissions

The assignment will be held as a competition. 50% of your grade will be based on how well you do in the competition (another 50% will be based on the novelty and soundness of your algorithm written in the report). The top team(s) for each task will receive “secret presents”.

You need to submit not only the results but also a report (in .doc, .txt or .pdf) that clearly describes your method. Note that you do not need to submit your code this time.

Deadline

2011/12/18(Sun) 23:59