

CS 395T Data Mining, Spring 2007: Project Proposal

David Chen, Xin Li

March 14, 2007

1 Background

We plan to work with the CS Journal data. This data contains the number of publications in various journals and conferences in Computer Science for different institutions during the past ten years. This data is interesting as it provides some information about the relation of different journals and institutions in Computer Science. We can try to do a variety of analysis on this data.

2 Plan of attack

We plan to do several analysis on the Journal data, including the following:

1. We would like to divide the data into training set and testing set, then, using regression methods, we can find a way to predict the number of publications of each institution in different journals in the future. The predictions can be tested on the testing set. To be more reasonable, we plan to divide the journals into different fields of Computer Science and do the regression and prediction in each field for different institutions.

2. We plan to do a co-clustering on this data set. As top institutions tend to publish more papers on top journals, a co-clustering of the institutions and journals will help us decide the ranking of institutions and journals. Again we plan to do this in different fields of Computer Science, in this way we can compare the rankings of institutions in different fields. This ranking estimation may also help us do the regression and prediction we proposed in 1.