1.File Structure

1.1 Raw Data

Contain untouched original data : XML files and txt files, storing contents of papers and user preferences, respectively.

Downloaded from : http://www.comp.nus.edu.sg/~sugiyama/SchPaperRecData.html

1.2 Processed Data

Contain the dictionary, feature matrix and filtered .mat file.

1.3 Java Codes

Contain the java program we wrote to parse the raw XML file, output dictionary and feature matrix.

1.4 Matlab Codes

Contain all learning algorithms we implemented, as well as a pre-process function that outputs DataX.mat and DataY.mat.

2.How to run

We did not borrow any external codes or software products.

2.1 The Pre-processing of data.

To reduce the size of the zip ball, we deleted redundant raw data from both Java and Matlab work directory, since all data have been fully pre-processed. But if you are interested in our data processing procedure, please copy files from the “Raw Data” directory to proper locations so our program could load it.

2.2 The major part

All enclosed Matlab codes are ready to run. We arranged Boosting and ANN separately in two working directories.

2.2.1 Boosting

The script “test.m” has everything you need to test the AdaBoost algorithm.

Scripts “pre_proc” are what we used to further filter the processed data.

All the other files are Matlab functions and cannot directly run.

2.2.2 ANN

Similarly, run “testann.m” to test the ANN algorithm we implemented. The “ann.m” was used to train the network and you do not need to re-run it.