Tracy Oguni

Data Mining

Project Progress Report for Honors Credit

April 2, 2015

# Current Status

The following explains the current status of the project:

## Project Objective

The objective is to collect rating information (including star ratings and comments) about professors at UT Arlington and put them into distinct groups based on their ratings and performance at the university. The professors will be sorted into categories such as “Interesting”, “Easy” or “Unknowledgeable”. Professors will be able to belong to more than one category. This project will help better understand the kind of professors found at UT Arlington and identify the common complaints made by students.

## Tasks

1. Downloading appropriate webpages from the websites
   1. Download the ratemyprofessor webpages for UT Arlington professors
   2. Download the UTA SFS webpages
2. Mining the web pages downloaded for rating information
   1. Convert the tables from the UTA SFS webpage into an internal data structure
   2. Convert the ratemyprofessor webpage information into an internal data structure
3. Cluster the data based on its attributes
4. Calculate the similarity between members of the same cluster and dissimilarities between members of different clusters.
5. Plot the data on a graph

## Deliverables

At the end of the semester, the data to be delivered will be:

1. A description of the various groups that professors can be divided into.
2. Graphs showing the distribution of professors into the various divisions.
3. A scatter plot graph showing the distance between individual professors on separate attributes.

This same data will be put on the project website and presented before the class in a demo.

## Challenges of the Project

A major challenge in the project was getting the information from the webpages. The UT Arlington SFS dataset is a simple table, but the RateMyProfessor data is more complex and required more time to accurately parse the download. Compared to this, clustering the data and dividing the professors into classes will be easy.

In addressing this, I have successfully downloaded several webpages and extracted the data I need from them. Doing this before the progress report means that the risk has been successfully handled.

## Methods and Algorithms

No special clustering algorithm will be used. The average of each teacher’s score in the relevant categories will be plotted on a graph and the teachers will be classified based on their similarity to each other. The categories to evaluated will be Clarity, Helpfulness, Easiness, and the six ratings defined in the UT Arlington SFS as follows:

1. The instructor provided clearly defined expectations (CDE).
2. The instructor used teaching methods that facilitated my learning (GTM).
3. The instructor encouraged me to take a role in my own learning, to ask questions, and to participate (SP).
4. The instructor was well prepared to teach (WP).
5. The instructor was available outside of class either electronically or in person (Available).
6. The instructor is one whom I would recommend to other students (Recommended).

## Initial Implementation

So far, I have successfully downloaded several ratemyprofessor.com webpages and extracted the information I need from them. I have also created an immutable java structure to store them. All work will be put on GitHub.

## Evaluation Plan

As previously said in the project proposal, evaluating the solution will be done by comparing the similarity and dissimilarity of professors inside and between clusters. Higher similarity inside clusters and higher dissimilarity between clusters will indicate a good clustering.

## Changes Since Project Proposal

There have been no changes made since the project proposal.

## Difficulties Encountered

See the section Challenges of the Project for information about difficulties encountered.

# Tasks to be Accomplished

The following describe the remaining tasks to be accomplished.

## What is left to be done

1. Download all the need ratemyprofessor.com webpages and extract the needed information from them.
2. Extract data from the UT Arlington SFS webpages.
3. Mining the web pages downloaded for rating information
   1. Convert the tables from the UTA SFS webpage into an internal data structure
   2. Convert the ratemyprofessor webpage information into an internal data structure
4. Cluster the data based on its attributes
5. Calculate the similarity between members of the same cluster and dissimilarities between members of different clusters.
6. Plot the data on a graph

## Finishing the Project

The project will be finished before the final report is due. A presentation will be prepared and the results delivered as per the instructions given by the professor

## Expected Challenges

One major challenge has revealed itself and is expected to occupy the student in the following days. There are a lot of ratemyprofessor.com webpages that must be manually downloaded unless some other way can be found. This will be very time consuming.