CS 410 Final Project Progress Report: Improving a System

Captain: Philip Cori (pcori2), Team Member 1: Henry Moss (htmoss2), Team Member 2: Kyle Maxwell (kylem6)

Collaborative Filtering with Social Exposure: A Modular Approach to Social Recommendation

Progress Made So Far

Our team has been able to recreate the data results from the paper above, using RecQ, a Python library for recommender systems which includes the SERec algorithm.

We have collected preliminary results on a particular approach for measuring the closeness between friends. We experimented with using the number of mutual friends as the driving factor in measuring closeness. Specifically, we used the formula:

Closeness = m / n

Where m is the number of mutual friends between the two friends and n is the total number of friends between the two friends. This seemed like a reasonable approach given that based on intuition, it seems two friends would be closer if they have more mutual friends. Not only this, but two friends are also considered closer if they have less total friends, which gives more weight to their own friendship.

All results use 5-fold cross validation.

Baseline results of SERec algorithm recreated on local machine:

- Precision: 0.0449469214437
- Recall: 0.455598102652
- F1: 0.0818217388808
- MAP: 0.146067212641
- NDCG: 0.334412555937

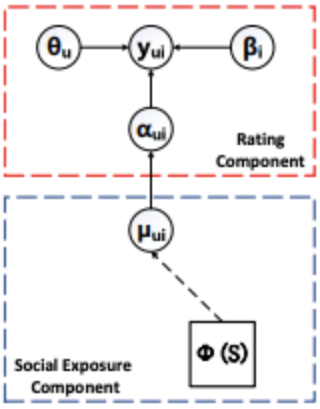
Results using mutual friends closeness measure:

- Precision: 0.0455850128798
- Recall: 0.461858940973
- F1: 0.0829799791008
- MAP: 0.149447844847
- NDCG: 0.340728707592

As can be seen, we've managed to increase the NDCG measure by 1.89% and the MAP measure by 2.31%.

The second improvement strategy we are exploring is the modification of the matrix factorization model in the Rating Component of the SERec Boost algorithm. Incorporating the Weighted Rating Matrix Factorization methods from Collaborative Filtering for Implicit Feedback Datasets (Yifan Hu et al, KDD 2009).

This matrix factorization model uses the implicit feedback data as an indication of positive and negative preference associated with vastly varying confidence levels. Primarily, we are layering this model to initialize the latent factors theta and beta, the user preferences and item attributes respectively. The social exposure component is then incorporated afterwards, only requiring a small number of iterations for the expectation-maximization algorithm to compute.



Results from modified matrix factorization:
Top 100, 5-fold cross validation

- Precision: 0.0493
- Recall: 0.4993
- F1: 0.0897
- MAP: 0.1656
- NDCG: 0.3702

This is a 10.7% increase in NDCG and 13.4% increase in MAP score from the baseline results.

Remaining Tasks

Regarding measuring closeness between friends, we will continue to experiment with other social network concepts in measuring closeness. The Lastfm dataset also contains data not currently used by the SERec algorithm, so we will also investigate whether this can be used as well. This data includes data about tags on different artists and which users placed which tags. We are also given time stamps to record when such tagging events occurred.

Regarding the matrix factorization, there is plenty of potential to improve the integration of the Weighted Rating Matrix Factorization model. As mentioned previously, we layered the components, but it seems that the Social Exposure Component could be incorporated in all iterations of the WRMF training.

Additionally, we will likely conduct tests of significance for our attempted improvements to tell whether they are actual improvements or due to randomness in the data.

Challenges/Issues Faced So Far

Understanding the architecture and class structure of the RecQ system has taken the majority of our work so far. It takes time to identify code areas that are adjustable that will still keep the system functioning. An additional challenge is the time needed to collect results. Running the 5 iterations of the EM algorithm proposed takes roughly 30 minutes to complete.

Furthermore, it has been challenging to improve the model based on social contagion, as suggested in the original paper. Social contagion is such a broad concept that it has been more difficult than expected to understand how to quantify and incorporate this idea into our preexisting model. We will continue looking into this as well as researching any possible improvements to be found from social structural influence, which was another method mentioned in the original published paper.