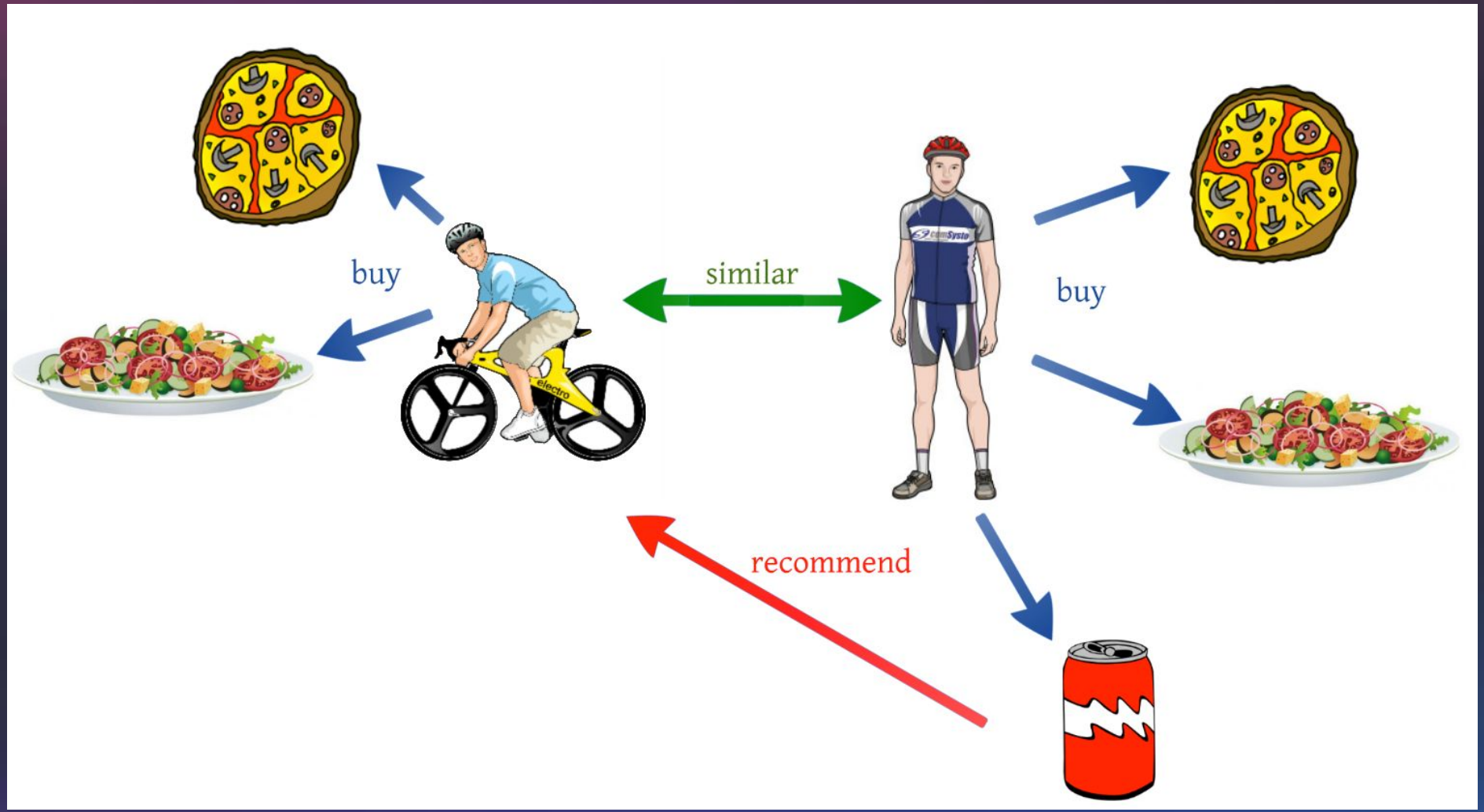
Algorithms: Group 6

David Arredondo, Sophie Beiers, Huijun Cui, Richard Shin, Shan Zhong

Collaborative Filtering



Collaborative Filtering

- An automated way to make predictions about *user A* that utilizes collective interests and preferences from similar users. Ultimately; able to recommend something *user A* might like.

this or… that?

this or... that?

Project: Datasets

Microsoft Web Data

- 4151 users, total of 33,875 vroots visited
- Binary 1’s, 0’s for each vroot

EachMovie Data

- 5055 users rated 1597 movies
- Ratings from 1-6

Project: Models

- Build two collaborative filtering algorithms:
  - Model-based clustering algorithm (EM algorithm)
  - Memory-based algorithm
- Consider different similarity weights for memory-based model:
  - Spearman, Pearson, vector similarity, entropy, mean-square difference and SimRank
  - Group 6 also tasked with rating normalization
- Compare prediction accuracies and trade-offs

Memory-based Model

- Generate prediction based on user A's similar neighbors' preferences and scores.
  - Calculate correlation/similarity weights between users and take weighted average of all users for prediction.

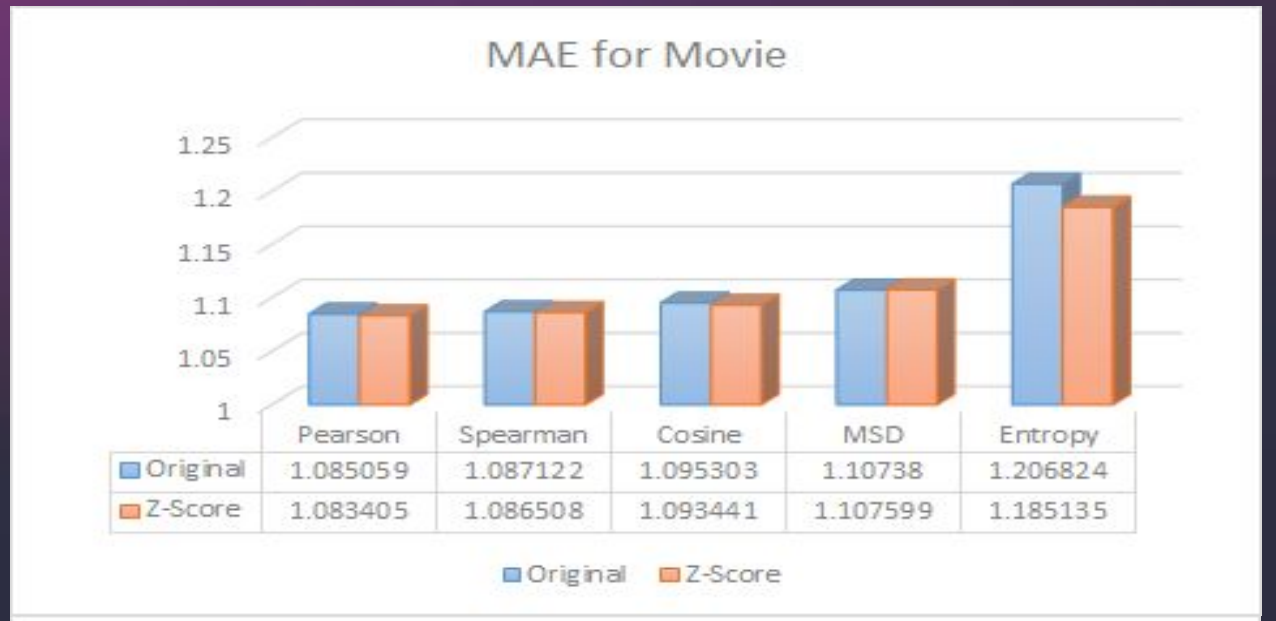
Pearson's Correlation:

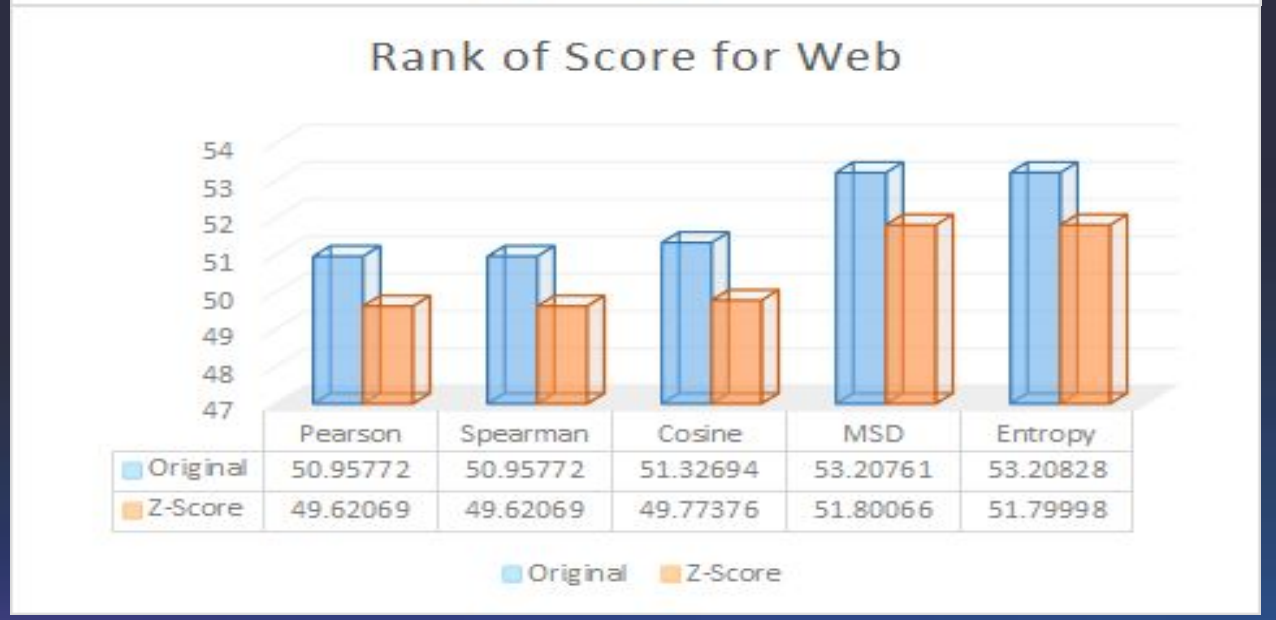
$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[\, n\sum x^2 - (\sum x)^2 \,][\, n\sum y^2 - (\sum y)^2 \,]}}$$

Memory-based Predictions

- MAE

- ROC





Model-based: EM Algorithm

Computation speed: Used a mixture of for loops and matrix operations; the three dimensions made matrix operations difficult

Starting point: Used a dirilecht distribution. The uniform distribution go stuck in a lower optima

Convergence: Trouble with likelihood; unsure if calculated correctly, because the algorithm would converge in ~10 iterations.

EM Predictions

- Wrote EM Predictor function based on:

$$\hat{r}_{a,m} = \mathbb{E}(R_{a,m} \mid r_{a,j}, j \in I(a)) = \sum_{k=1}^{5} k \, \frac{\sum_{c=1}^{C} \mu_c \gamma_{m,c}^{k} \prod_{j \in I(a)} \gamma_{j,c}^{r_{a,j}}}{\sum_{c=1}^{C} \mu_c \prod_{j \in I(a)} \gamma_{j,c}^{r_{a,j}}}.$$

- Idea: The prediction is a weighted avg. of the rating values weighted by the gamma probabilities of being in that cluster.
- Took FOREVER