# A Personalized Recommendation System for Yelp Users

Yinuo Yao[1], Xin Niu[2], Fangmingyu Yang[2]

[1] Civil and Environmental Engineering, Stanford University; [2] Management Science and Engineering, Stanford University

## Introduction

Yelp, a platform that consists of various business information, allows the users to provide star ratings and text reviews for businesses they visited before. This provides insights for other potential users. Particularly for restaurants, most users choose based on two different considerations, the overall ratings and the reviews posted by other users.

However, there is a severe problem associated with this approach. The current rating system only provides an average value without considering any personalized information of the individual user. Thus, the efficacy of the rating system is diminished severely. It is not uncommon for the user to think of a restaurant as overrated or underrated after visiting. The underlying cause of this problem is the inability to provide a personalized rating.

## Objective

The objective of this project is to construct a more sophisticated model that provides a personalized star ratings of the restaurants based on more features such as the similarity among different users.

## Methodology

### General Outline of Approach

- Need to extract the group of users or restaurants relevant to the query we receive.
- But data does not contain relevant users' preferences.
- Information can be extracted from the dataset into two similarity metrics (User-User & User-Category)
- Given a query about user U and business B, in pseudocode format, we can describe the algorithm as follows,

**Algorithm 1: Determining Personalized Rating**

```
1   Personalized Rating (U, B)
2   {
3       GET categories of B as C
4       CALCULATE similarity (U, C)
5       FOR each User
6           CALCULATE similarity (U, User)
7           IF User rated B
8               ADD User to Pool
9       CALCULATE PredictedRating of B by Pool
10      RETURN PredictedRating
11  }
```

## Methodology (Continue)

### Similarity Metric (User-User)

- Based on the heuristic method and the weighted Jaccard similarity, which defines a metric $0 \le si\ m(u, i) \le 1$.

**Step1:** Calculate the number of categories $B_i$ has.

$$m_{C_p} = \sum_{i=1}^{m} 1\{B_i \in C_p\}$$

**Step 2:** Calculate the percentage of restaurants that user $i$ and user $j$ both went and gave same ratings.

$$sim(u_i, u_j) = \frac{\sum_{k=1}^{m_{C_p}} 1\{\bar{s}_i^{(k)} = \bar{s}_j^{(k)} \wedge v_i^{(k)} = v_j^{(k)}\}}{\sum_{k=1}^{m_{C_p}} 1\{v_i^{(k)} = v_j^{(k)}\}}$$

### Similarity Metric (User-Category)

- Measured the similarity between the individual user's past reviews and the restaurant.

**Step 1:** Sum up the number of restaurants belonging to category $C_p$ as a fraction of the total number of restaurants user $i$ visited

$$sim(u_i, C_p) = \frac{\sum_{k=1}^{N_R^{(i)}} 1\{B_k \in C_p\}}{N_R^{(i)}}$$

### Algorithm

**Step 1:** Calculate the centered rating of restaurant $b_k$ by user $u_j$ for category $C_p$

$$d_j = s_j^{(k)} - \frac{\sum_{l \neq k}^{m_{C_p}} v_j^{(l)} s_j^{(l)}}{\sum_{l \neq k}^{m_{C_p}} v_j^{(l)}}$$

**Step 2:** Aggregate the centered ratings using User-User Similarity Metric and adjust to scale of user $u_i$

$$\widehat{s_{C_p}}_i^{(k)} = \frac{\sum_j sim(u_i, u_j) d_j}{\sum_j sim(u_i, u_j)} + \frac{\sum_{l \neq k}^{m_{C_p}} v_i^{(l)} s_i^{(l)}}{\sum_{l \neq k}^{m_{C_p}} v_i^{(l)}}$$

**Step 3:** Calculate the final predicted rating by combining using User-Category Similarity Metric

$$\hat{s}_i^{(k)} = \frac{\sum_{C_p} sim(u_i, C_p) \widehat{s_{C_p}}_i^{(k)}}{\sum_{C_p} sim(u_i, C_p)}$$

### Error Rate Calculation

$$Error\ Rate = \frac{\sum |predicted - actual|}{\sum actual}$$

### Notation

Where $\sum_{P=1}^{N_c} m_{C_p} \ge m$ since one business ID can have multiple labels, $v_i^{(k)}$ is 1 if user $i$ has visited business $k$, $\bar{s}_i^{(k)}$ is the average star ratings on business ID k from user $i$, $B_i$ is the $i^{th}$ business ID, m is the number of all the unique business ID drawn from the business data set, $m_{C_p}$ is the number of business ID which belongs to the category $C_p$, $N_c$ is the total number of categories associated with the restaurant business type, $N_R^{(i)}$ is the total number of review from user $i$.

## Results and Discussion

### Results of Similarity Metrics

- Figure 1 and Figure 2 demonstrated User-User Similarity Metric and User-Category Similarity Metric respectively with a small set data.

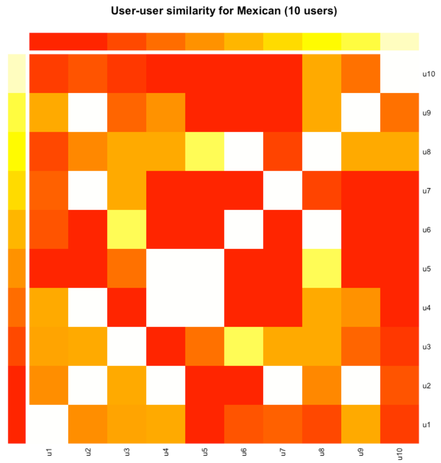


**FIGURE 1:** Example of a User-User Similarity Metric presented in the form of Heat map (10 Users for Mexican Restaurant)

- The diagonal of Figure 1 has the highest value because the each person is most similar to himself.

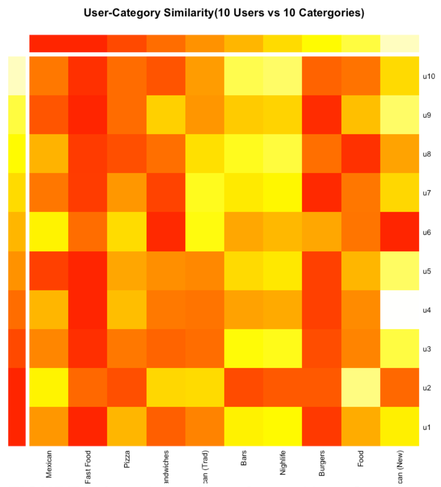


**FIGURE 2:** Example of a User-Category Similarity Metric presented in the form of Heat map (10 Users and 10 Top Categories)

### Rating Prediction Results

- Figure 3 demonstrated the comparison between the calculated ratings using the algorithm we proposed and the actual rating the users give to the restaurant.
- As shown in the Figure 3, Predicted rating (Orange) was very close to the actual rating (Green).
- Figure 4 demonstrated the comparison between the calculated ratings and actual ratings given by an user to all the restaurants he visited.
- Predicted rating (Orange) lay between Yelp rating and User rating which indicated an improvement in the personalized rating.

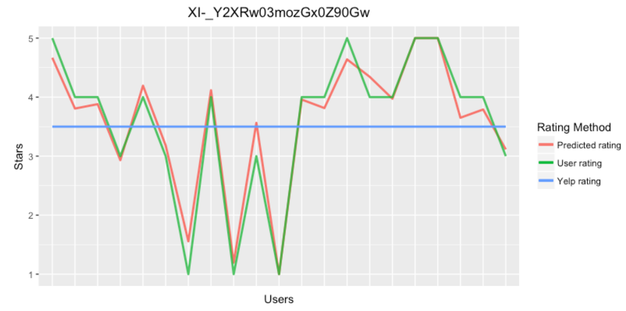


**FIGURE 3:** Comparison of Predicted Rating, User Rating & Yelp Rating

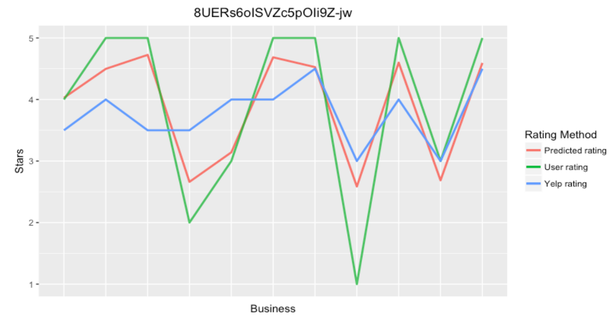


**FIGURE 4:** Comparison of Predicted Rating, User Rating & Yelp Rating

- Table 1 demonstrated the error rate of prediction as compared to the actual rating given by the users for 5 different restaurants.

**TABLE 1:** Error Rate of Prediction Produced by Algorithm

| Restaurant Yelp ID | Error Rate |
|---|---|
| **lf2DUhmWlvlu2JFc_rR-Bw** | 0.0950 |
| **yExYqENb4F6qH6kJxTOaSQ** | 0.1028 |
| **LXhL5X3edNRy7epku6UAEw** | 0.0883 |
| **XI-_Y2XRw03mozGx0Z90GW** | 0.0265 |
| **kR5i58Pcse1FD9tk-yYLIA** | 0.0807 |

## Conclusions & Future Work

- Using Similarity Metric, both User-User and User-Category, indeed provided a better personalized rating system.
- The predicted rating was much closer to the actual rating given by the user than the average rating provided by Yelp.
- **Future:** Extend the study to the text part of the review in addition to the rating