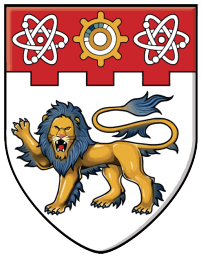


NANYANG
TECHNOLOGICAL
UNIVERSITY

Your Neighbors Affect Your Ratings: On Geographical Neighborhood Influence to Rating Prediction

Longke Hu **Aixin Sun** Yong Liu

Nanyang Technological University

Singapore

Outline

The problem: business rating prediction

Rating prediction is to predict the preference rating of a user to a product or service (*i.e.*, an item) that she has not rated before.

- A well defined research problem in recommender systems
- An array of widely studied solutions, *e.g.,* collaborative filtering
- Users $\Longleftrightarrow$ Items: songs, movies, books. . .

A business is an item in our problem setting

- A business can be a restaurant, shopping mall, beauty salon . . .
- A business **physically exists at a specific geo-location** with latitude/longitude coordinates
- Most businesses are not geographically isolated from others

A business physically exists at a geo-location

When a user visits a business, there is a good chance that:

- She **walks by its neighbors** if they are located within walking distance.
- The **overall environment** of that region might affect her rating to the business.

Questions

1. Is it true that most businesses have neighbors in walking distance?
2. Is there any correlation between a business's rating and its neighbors' average rating?
3. Is the category of a business a factor here?

The Yelp dataset

Was used in ACM RecSys Challenge 2013

- Sampled from the greater Phoenix, AZ metropolitan area from March 2005 to January 2013
- 11,537 businesses, 229,907 reviews by 43,873 users, and 8,282 check-in sets

More details

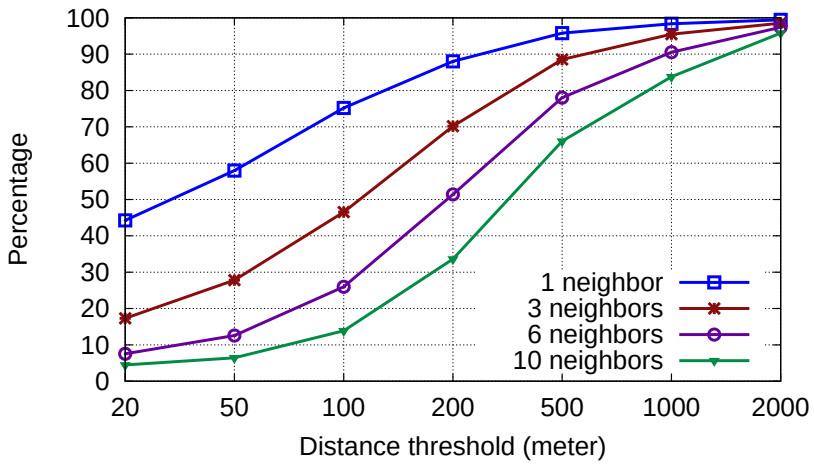
- A **business** has id, name, latitude longitude, categories...
- A **review** contains business id, user id, rating from 1 to 5 stars, date, review text, and voting.
- A **check-in set** for a business contains the aggregated number of check-ins in every hour from Monday to Sunday.

Geographical neighbors within walking distance?

Observation 1

Most businesses have **neighbors within a short geographical distance** from their locations.

Percentage of businesses having at least 1, 3, 6, 10 neighbors within a distance of 20 - 2000 meters.



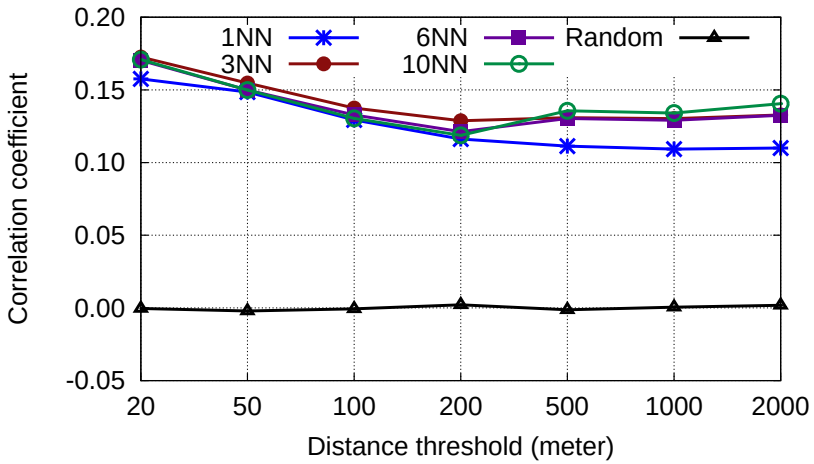
- More than 44% of businesses have one neighbor next to it within 20 meters.
- About 95% of businesses have one neighbor within 500 meters.

Business rating correlation?

Observation 2

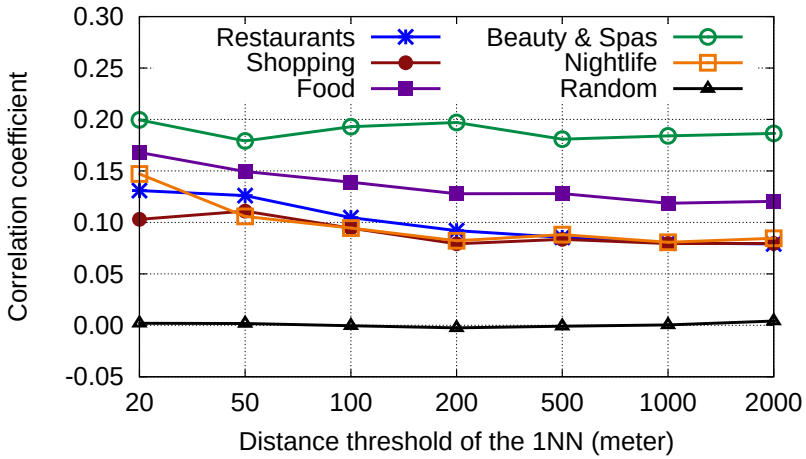
The average rating of a business is **weakly positively correlated** with the average rating of its neighbors.

Pearson's correlation coefficient between a business's rating and the average rating of its 1, 3, 6, and 10 nearest neighbors, at different distance thresholds from 20 to 2000 meters.

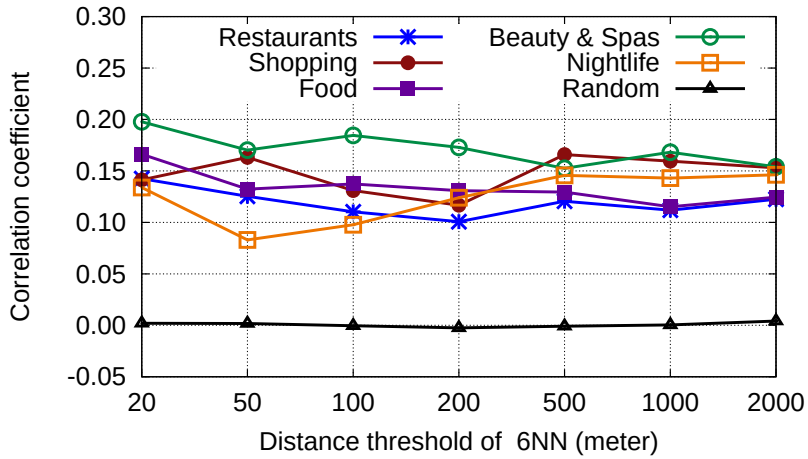


- Pearson's correlation coefficient is in the range of 0.109 to 0.173.
- The correlation is relatively stronger within a smaller distance.

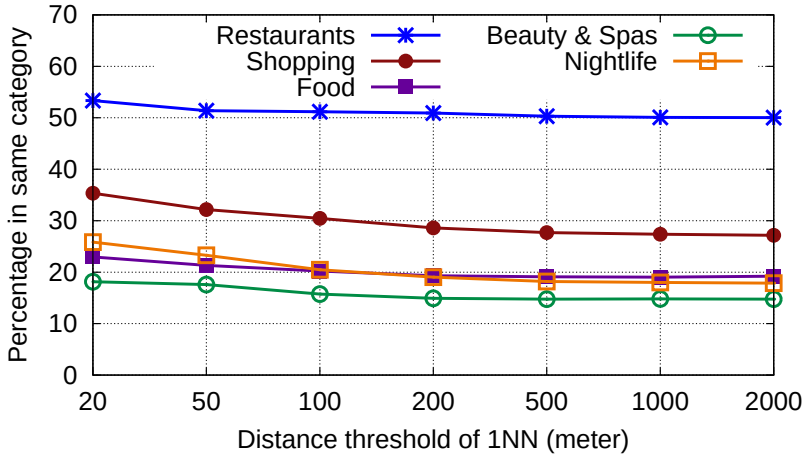
Is business category a factor?



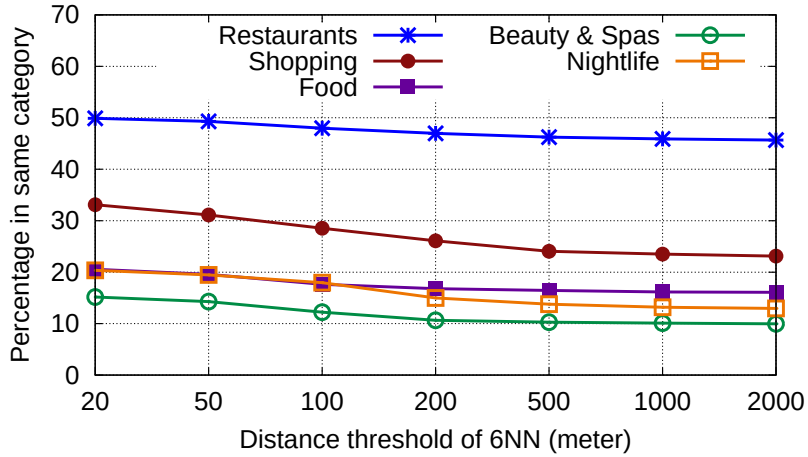
(a) Rating correlation of 1NN



(b) Rating correlation of 6NN



(c) % 1NN in same category



(d) % 6NN in same category

Questions and observations

1. Is it true that most businesses have neighbors in walking distance?

   Observation 1: Most businesses have **neighbors within a short geographical distance** from their locations.

2. Is there any correlation between a business's rating and its neighbors' rating?

   Observation 2: The average rating of a business is **weakly positively correlated** with the average rating of its neighbors.

3. Is the category of a business a factor here?

   Observation 3: The weak positive correlation in ratings is **independent of the categories** of the businesses and/or their neighbors.

Data analysis: a summary

Intrinsic characteristics

The rating of a business should **mainly** depend on the characteristics of the business itself, *e.g.,* quality of products or services, not its neighbors.

Extrinsic characteristics

"Things of one kind come together": A business is **not geographically independent** from its neighbors. These neighbors give a user the sense of the surrounding environment of the business, *e.g.,* hygiene standard.

Business rating prediction

Both the intrinsic and extrinsic characteristics of a business shall be modeled in rating prediction.

Related work: collaborative filtering

Collaborative Filtering: Similar users rate items similarly or similar items receive similar ratings from users.

Memory-Based CF

- Finding similar users or items by using similarity measures
- UserKNN, ItemKNN, Pearson's Correlation, Cosine similarity
- Similar users or items are also known as "neighbors"

Model-Based CF

- Building models from the observed user-item ratings
- Latent factor model: users and items are jointly mapped into a shared latent space of low dimensionality
- Matrix factorization models: Biased MF, SVD++, Social MF ...
- Evaluated on: Yahoo! Music, Last.fm, Netflix, Douban ...

Related work: POI recommendation and prediction

POI recommendation is to recommend unvisited POIs to users

- Geographical influence: Users tend to visit nearby POIs of their home/office locations; nearby locations of the POIs in their favor
- Temporal influence: Users check-in different types of POIs at different time slots of a day
- Social influence among friends

POI prediction is to predict which POI a user would visit next

- Based on user's current location/time, predict next POI to visit
- Both geographical and temporal influence have been considered.

Neighborhood influence: key differences

- User's point of view vs business's point of view
- User's cost of travel (time, monetary)

Business rating prediction: Biased Matrix Factorization

The basic idea of Biased MF

- Each user and each item is represented by latent factors $\mathbf{p}_u$ and $\mathbf{q}_i$
- The predicted rating $\hat{r}_{ui}$ is the inner product of the two, with biases

$$\hat{r}_{ui} = \mu + b_u + b_i + \mathbf{p}_u^\top \mathbf{q}_i$$

Parameter estimation

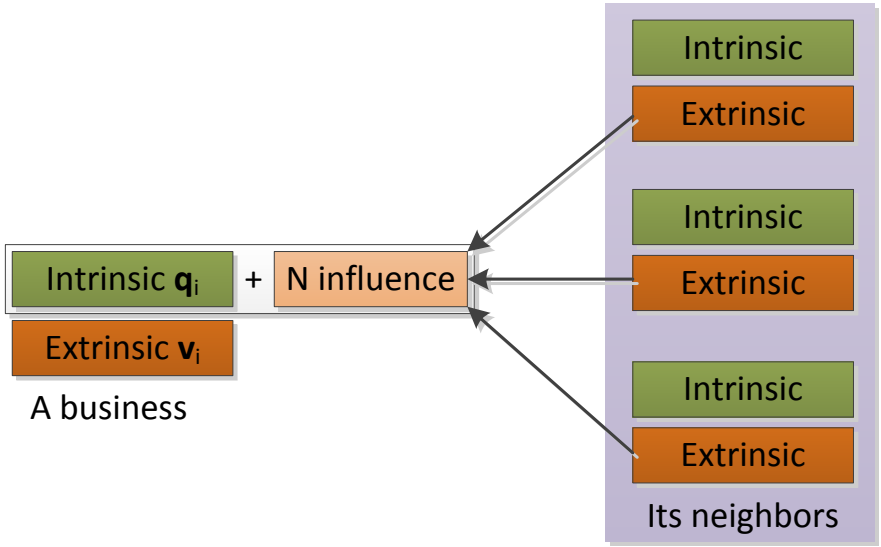
- Optimization: minimize regularized squared error on $\mathcal{K}$
- Algorithm: Stochastic gradient descent (SGD) and alternating least squares (ALS)

$$\min_{\mathbf{p}_*, \mathbf{q}_*, b_*} \sum_{(u,i)\in\mathcal{K}} (r_{ui} - \hat{r}_{ui})^2 + \lambda_1 \left(\|\mathbf{p}_u\|^2 + \|\mathbf{q}_i\|^2 \right) + \lambda_2 \left(b_u^2 + b_i^2 \right)$$

Incorporating neighborhood influence

Two kinds of factors of a business

- Intrinsic characteristics: latent factors $\mathbf{q}_i$
- Extrinsic characteristics: latent factors $\mathbf{v}_i$



With influence from neighborhood, the predicted rating $\hat{r}_{ui}$ is:

$$\hat{r}_{ui} = \mu + b_u + b_i + \mathbf{p}_u^\top \left(\mathbf{q}_i + \frac{\alpha_1}{|N_i|} \sum_{n \in N_i} \mathbf{v}_n \right)$$

Objective function is updated with regularization components for $\mathbf{v}_n$.

Incorporating category influence

Why category influence?

- Category of a business reflects the characteristics of a business
- Users may use different criteria in different categories
- POI recommendation achieves better accuracy by considering the categories of the POIs

Approach: Each category is modeled by a latent factors vector $\mathbf{d}_c$.

$$\hat{r}_{ui} = \mu + b_u + b_i + \mathbf{p}_u^\top \left(\mathbf{q}_i + \frac{\alpha_1}{|N_i|} \sum_{n \in N_i} \mathbf{v}_n + \frac{\alpha_2}{|C_i|} \sum_{c \in C_i} \mathbf{d}_c \right)$$

The objective function is updated with regularization components for $\mathbf{d}_c$

Incorporating review content

A user rating usually comes with a **textual review**

- Review elaborates the reason behind the rating
- Partially reflects the characteristics of the business

Approach:

- Map the review words to the same latent factors space.
- Decompose $\mathbf{q}_i$ into a combination of latent factors of review words

$$\text{business latent facotrs } \mathbf{q}_i \Rightarrow \frac{1}{|R_i|}\sum_{w\in R_i}\mathbf{q}_w$$

$$\hat{r}_{ui} = \mu + b_u + b_i + \mathbf{p}_u^\top \left(\frac{1}{|R_i|}\sum_{w\in R_i}\mathbf{q}_w + \frac{\alpha_1}{|N_i|}\sum_{n\in N_i}\mathbf{v}_n + \frac{\alpha_2}{|C_i|}\sum_{c\in C_i}\mathbf{d}_c \right)$$

Popularity and geo-distance influences

Both are distinctive features in POI recommendation

- Businesses in downtown area likely receive more visits
- Users tend to visit nearby POIs

Approach: model region popularity and geo-distance as biases

- Business popularity ρ_i : Number of reviews + number of check-ins
- Geo-distance $\tau_{u,i}$: Estimate a user's 'home location' by recursive grid search algorithm, then compute the distance to business

Rating bias z with two parameters β_i and β_u: $z = \beta_i \rho_i + \beta_u \tau_{u,i}$

$$\hat{r}_{ui} = \mu + b_u + b_i + z + \mathbf{p}_u^\top \left(\frac{1}{|R_i|} \sum_{w \in R_i} \mathbf{q}_w + \frac{\alpha_1}{|N_i|} \sum_{n \in N_i} \mathbf{v}_n + \frac{\alpha_2}{|C_i|} \sum_{c \in C_i} \mathbf{d}_c \right)$$

Five factors in business rating prediction

- Neighborhood influence
- Category influence
- Review content
- Popularity bias
- Geo-distance bias

$$\hat{r}_{ui} = \mu + b_u + b_i + z + \mathbf{p}_u^\top \left(\frac{1}{|R_i|} \sum_{w \in R_i} \mathbf{q}_w + \frac{\alpha_1}{|N_i|} \sum_{n \in N_i} \mathbf{v}_n + \frac{\alpha_2}{|C_i|} \sum_{c \in C_i} \mathbf{d}_c \right)$$

$$z = \beta_i \rho_i + \beta_u \tau_{u,i}$$

Experiment setting

Yelp dataset

- Removal of businesses and users having fewer than 10 reviews
- Stopword removal and stemming in reviews
- 113,514 ratings by 3,965 users to 3,760 businesses
- For each user, 70% ratings used for training, 30% for testing

Evaluation metric

- Mean Absolute Error: $MAE = \frac{1}{|\mathcal{T}|} \sum_{(u,i)\in\mathcal{T}} |r_{ui} - \hat{r}_{ui}|$
- Root Mean Square Error: $RMSE = \sqrt{\frac{1}{|\mathcal{T}|} \sum_{(u,i)\in\mathcal{T}} (r_{ui} - \hat{r}_{ui})^2}$

Experimental results

Method	MAE	RMSE
Global Mean (μ)	0.8854	1.0962
Item Mean	0.8369	1.0939
User Mean	0.8599	1.0838
Item KNN	0.8208	1.0574
User KNN	0.8110	1.0429
Biased MF	0.8237	1.0483
SVD++	0.8120	1.0352
Social MF	0.8123	1.0303
N-MF	0.7952	1.0110
NC-MF	0.7929	1.0096
NCR-MF	0.7923	1.0078
NCRP-MF	**0.7920**	**1.0072**
NCRPD-MF	0.7958	1.0132
CRP-MF	0.7956	1.0138
CRPD-MF	0.8062	1.0191

Method comparison

- 8 baseline methods
- 7 proposed methods

N	Neighborhood influence
C	Category influence
R	Review content
P	Popularity bias
D	Distance bias

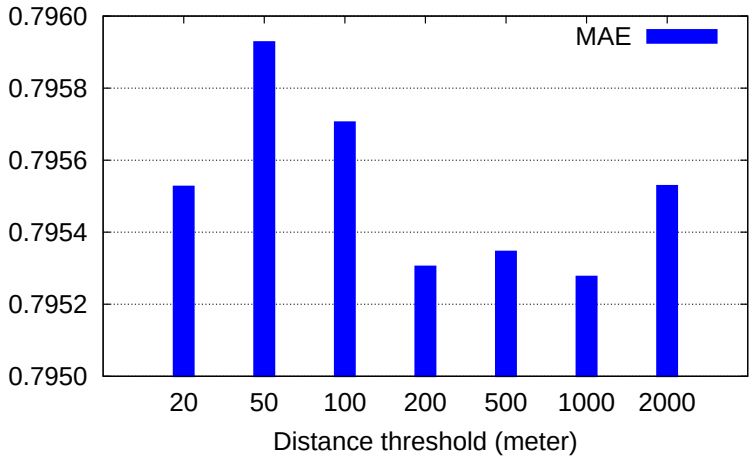
Experimental results: observations

1. Methods with **geographical neighborhood influence** outperform all baseline methods

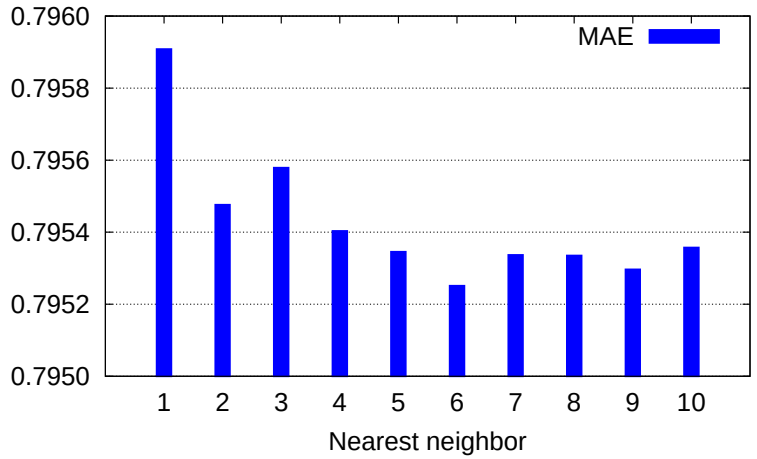
2. The best prediction accuracy is achieved by NCRP-MF; NCRPD-MF is poorer than N-MF
   - ✓ Geographical neighborhood (**N**)
   - ✓ Business category (**C**)
   - ✓ Review content (**R**)
   - ✓ Business popularity (**P**)
   - × Geo-distance (**D**)

3. SVD++, Social MF, and User KNN are the three best methods among baselines

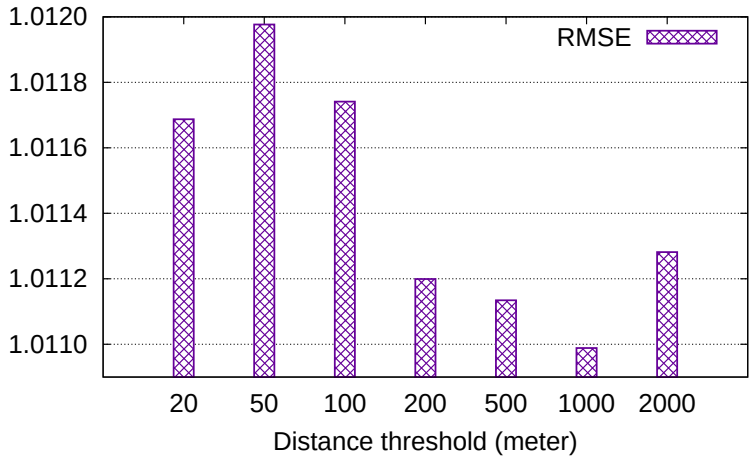
Impact of neighborhood size



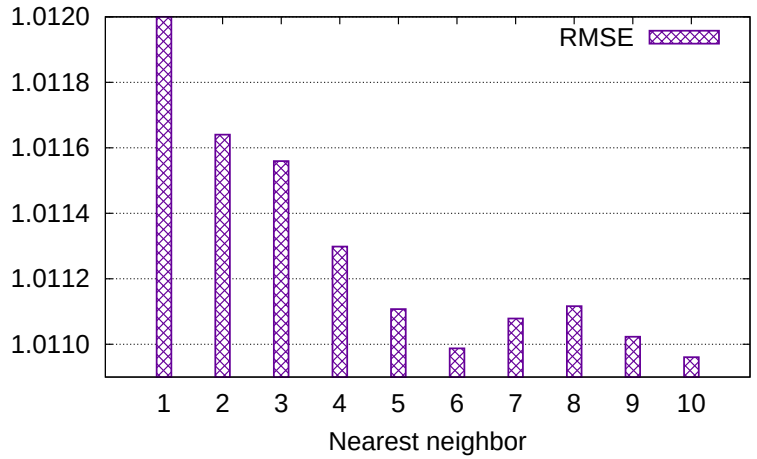
(e) Neighbors by distance (MAE)



(f) By neighborhood size (MAE)



(g) Neighbors by distance (RMSE)



(h) By neighborhood size (RMSE)

Cold-start business rating prediction

Predict ratings of existing users to "new" businesses

- Users: appear in our training data ($\mathbf{p}_u$ and b_u are known)
- Businesses: removed in data pre-processing for having fewer than 10 reviews ($\mathbf{q}_i$ and b_i are unknown)
- 20,395 ratings made by 3,319 existing users to 6,939 "new businesses"

Known factors:

- Global mean μ
- User mean μ_u
- User latent factors $\mathbf{p}_u$
- User bias b_u
- Neighbor latent factors $\mathbf{v}_n$
- Category latent factors $\mathbf{d}_c$

Method	MAE	RMSE
Global Mean	1.0319	1.2749
User Mean	0.9963	1.2566
Biased MF	1.0020	1.2539
N-MF	0.9956	1.2538
NC-MF	**0.9936**	**1.2535**

Conclusion

1. A business has a **physical location** and a business **has neighbors**.
2. A business's rating is **weakly positively correlated** with its geographical neighbors' rating.
3. We extend the Biased MF model to include both **intrinsic characteristics** and **extrinsic characteristics** of a business.
4. We show that geographical **neighborhood influence**, business category, popularity, and review content improve rating prediction accuracy.
5. We show that geographical distance between a user and a business adversely affects the prediction accuracy.

Which neighbors to consider?

NANYANG TECHNOLOGICAL UNIVERSITY

Dr. Aixin SUN
axsun@ntu.edu.sg
http://www.ntu.edu.sg/home/axsun/