User Check-In History Based Venue Recommendation

Boya Yu Jiamin Xuan Yuxiang Zhang New York University, CUSP

Contribution

We present an application of collaborative filtering for the venue recommendation in New York City. In the first part, we implement traditional memory-based collaborative filtering method to recommend venues when we train on user check-in history. Next, we combine with user relationship or clustering labels as weighting function that surpasses traditional memory based collaborative filtering method.

Data

- **Check-In Dataset**

We get data indirectly from Four-square check-in records shared on Twitter. The data were a sample of record from Feb. 2014 to Feb. 2015, parsed using PySpark from 113 GB geotagged tweets around NYC. Next we used the Foursquare Venue Search API to scrape venue information denoted by latitude and longitude. After preprocessing, we got 838 venues in which 4,365 users have checked-in 54,017 times in total.

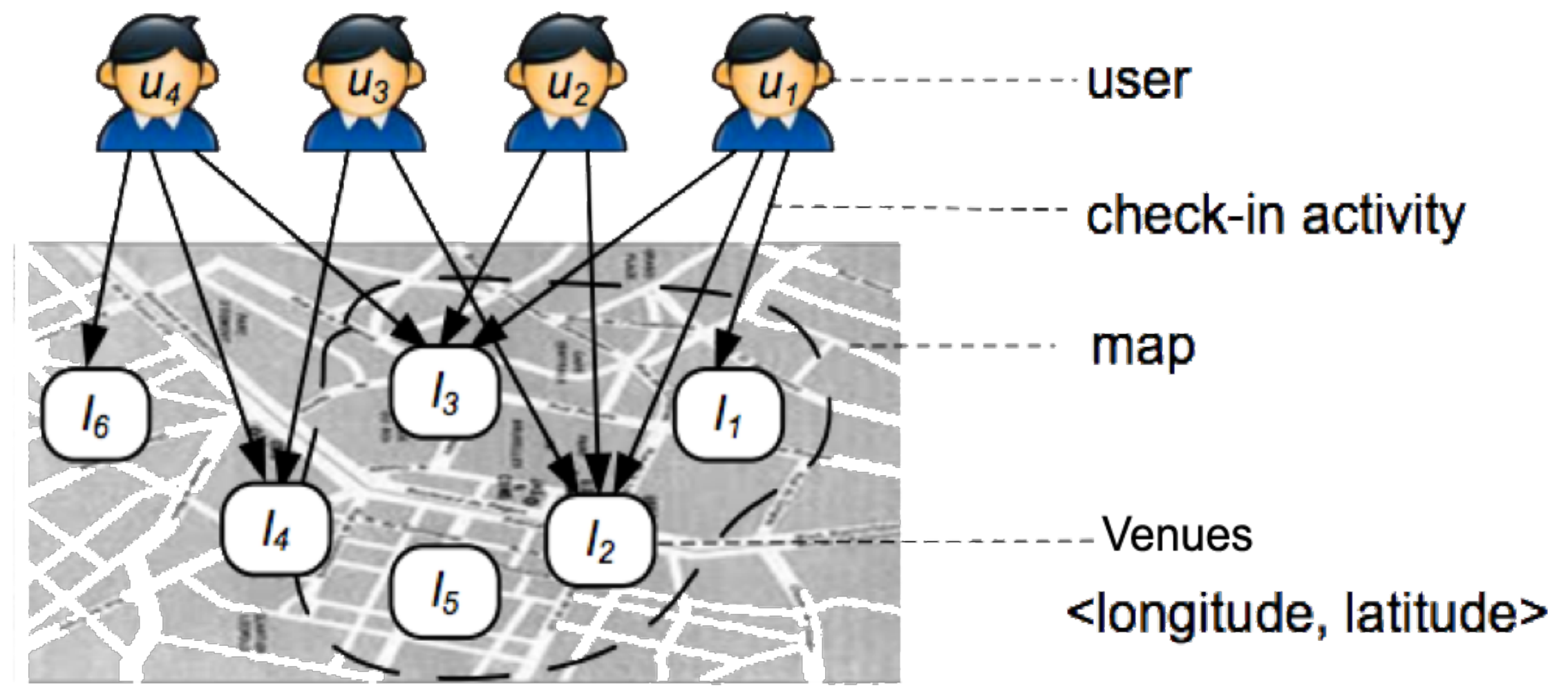
- **Venue Similarity**

For all the venues, we use Cosine Similarity and Jaccard Distance Similarity to transform the sparse matrix of user-venue check-in history into venue similarity scores.

- **User Relationship**

From a social media user's standpoint, following certain users may indicate a tendency of referring to their online activities. We scraped all the Twitter IDs that each of our users follows and discarded those IDs which are not in our set of 4,365 users.

Memory-Based CF



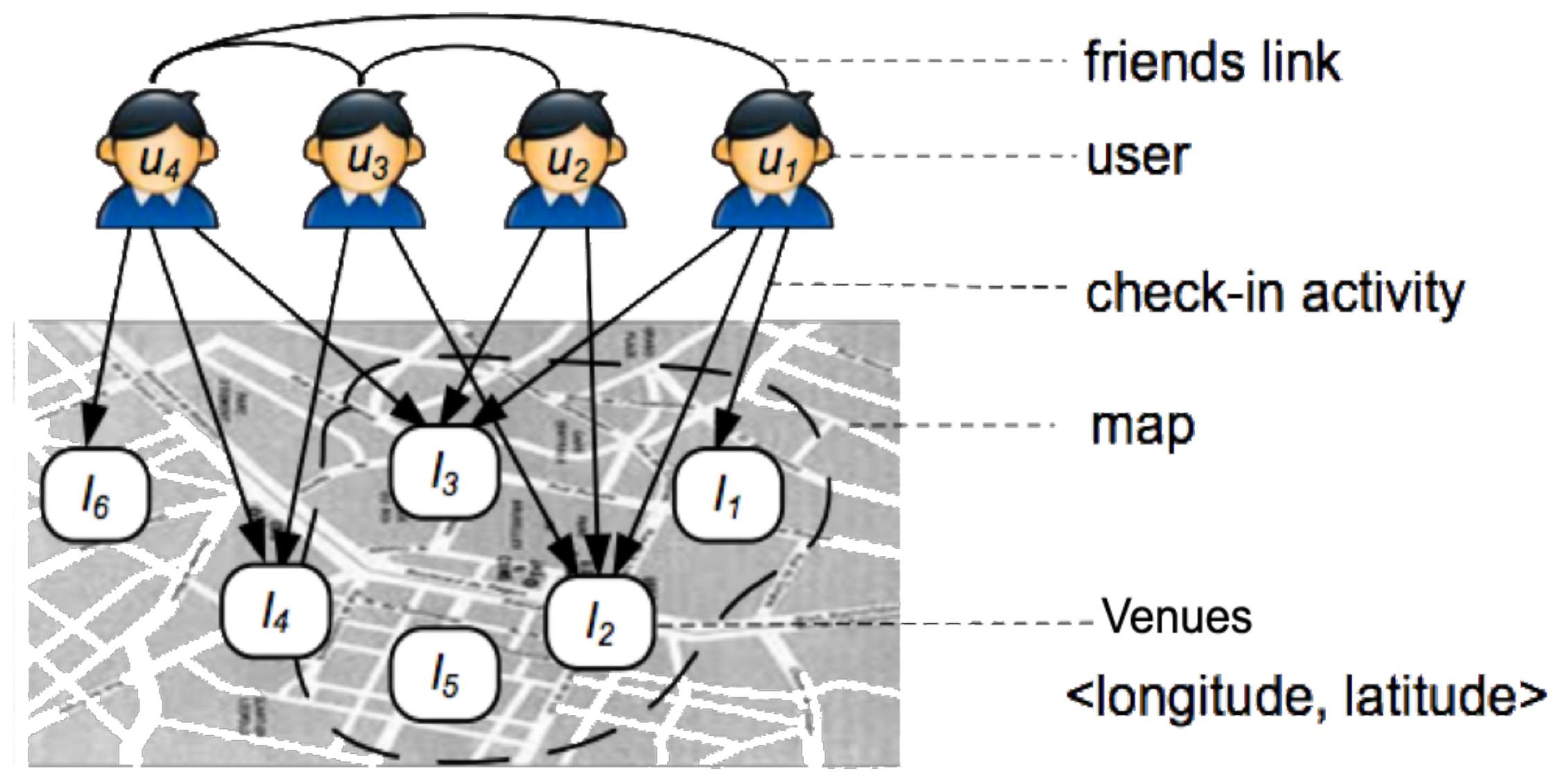
$$\hat{c_{i,j}} = \sum_{u_{k \in V}} w_{i,k} \times c_{k,j}$$

Where, $\hat{c}_{i,j}$: $Probability\ that\ user\ u_i\ has\ a\ check\ in\ at\ a\ new\ venue\ v_j$

$c_{k,j}$: $Indicating\ if\ user\ u_k\ had\ a\ check\ in\ at\ venue\ v_j$

$w_{i,k}$: $The\ similarity\ (cosine\ distance)\ between\ user\ u_i\ and\ u_k$

Trust-Based CF



$$\tilde{c_{i,j}} = \sum_{u_{k \in V}} (w_{i,k} + \alpha \hat{w_{i,k}} + \beta \tilde{w_{i,k}}) \times c_{k,j}$$

Where, $\tilde{c_{i,j}}$: $Probability\ that\ user\ u_i\ has\ a\ check\ in\ at\ a\ new\ venue\ v_j$

$c_{k,j}$: $Indicating\ if\ user\ u_k\ had\ a\ check\ in\ at\ venue\ v_j$

$w_{i,k}$: $The\ similarity\ (cosine\ distance)\ between\ user\ u_i\ and\ u_k$

$\hat{w_{i,k}}$: $The\ relationship\ (0\ or\ 1)\ between\ user\ u_i\ and\ u_k$

$\tilde{w_{i,k}}$: $The\ same - cluster\ indicator\ between\ user\ u_i\ and\ u_k$

Spectral Clustering and Parameter selection

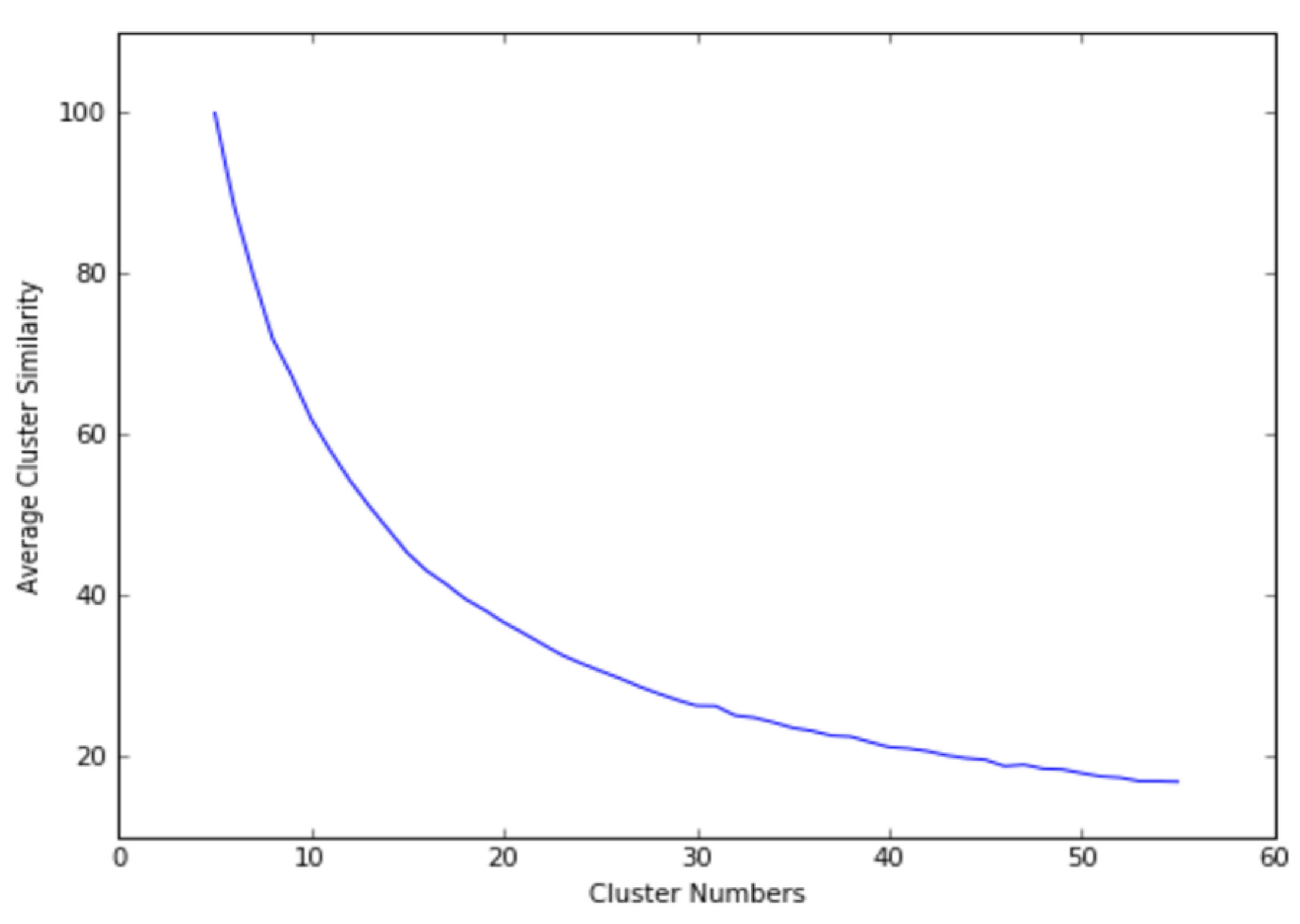


Figure: Average Cluster Similarity vs Number of Clusters

$$Average\ Cluster\ Similarity = \frac{1}{n}\sum_{j=1}^{N}\frac{1}{n_j}[\sum_{p,q \in c_j} w_{p,q}]$$

Based on user similarity matrix, we can view the set of user as a connected graph. Every user node is connected with others according to their similarity, and if we divide the graph into different parts, a better division tends to have higher in-cluster similarity among all. We deployed spectral clustering [6] to separate users into subgroups. Average cluster similarity is used to compare different clustering models.

Results

With both recommendation approaches, for every one of 4,365 users, we got the probabilities of visits on each of the 838 venues said person has never been. After ranking venue probabilities for each user, we got a Top 10 list of potential venues that this user would be recommended. For example, one user has been to Thai Son (Vietnamese Rest.) 1 time, Junior's Restaurant (American Rest.) 1 time, MSG (Stadium) 1 time, Ellen's Stardust (Diner) 2 times, and Bubba Shrimp Co. (Cajun Rest.) 2 times, we will provide the recommendation as following,

Place Recommended	Category	Predicted Check-In %
Barclays Center	Basketball Stadium	1
Yankee Stadium	Baseball Stadium	0.077539078
Central Park	Park	0.067403303
The Metropolitan Museum of Art	Art Museum	0.034893101
Disney Store	Toy / Game Store	0.033414733
Columbus Circle	Plaza	0.027634547
Standard Deviation		0.339192915

Memory-Based Venue Recommendation Sample

Place Recommended	Category	Predicted Check-In %
Barclays Center	Basketball Stadium	1
Yankee Stadium	Baseball Stadium	0.869281718
Central Park	Park	0.450006654
The Metropolitan Museum of Art	Art Museum	0.430940557
Disney Store	Toy / Game Store	0.35639509
Columbus Circle	Plaza	0.35569186
Standard Deviation		0.269303363

Trust-Based Venue Recommendation Sample

Next Steps

There are a few areas worth exploring on top of what we have accomplished. First, there are certain features in our data that we did not incorporate into the analysis, such as check-in time, venue category, and user bio. Furthermore, matrix factorization is an arguably more sophisticated approach than memory-based and trust-based CF, because it ideally allows us to discover the latent features underlying the interaction between users and venues.

Acknowledgments

1. Ye, Mao, et al. "Exploiting geographical influence for collaborative point-of-interest recommendation." *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, 2011, pp. 325-334.
2. Luxburg, Ulrike von. "A Tutorial on Spectral Clustering." *Springer*, 17 (4), 2007.