Shi Fan, Tianyi Gu, Xiaoge Wu, Yuxiang Zhang
Professor Sobolevsky, Jankowiak, Shroff
Machine Learning for Cities, Spring 2016, CUSP

Time 2 Out New York with Machine Learning

May 1, 2016

Introduction

The dynamics of a city have been shaped by a wide variety of features including cultural perceptions, economics factors, demography, geography and resources [1]. In order to better understand local urban life, a new methodology for studying the dynamics, structure, and character of a city has been recognized on a large scale, which is fundamentally data-driven. Given geospatial social media data from thousands of users, we intended to develop a venue recommendation system to help New Yorkers find the best places to visit in New York City.

With the enormous amount of reviews and ratings available online, it is unpractical and inefficient for social media users to glance all types of information to decide where to go when it comes to exploring a city. With machine learning, we can approach the problem technically and give personalized recommendations from multiple perspectives. E-commerce companies usually recommend products based on similarity among users, while we sometimes rely on recommendations from our friends and people we follow [2]. In this project, we intended to integrate the two types of approach—memory-based and trust-based—together.

In order to develop such a venue recommendation system for New Yorkers, our group focused on utilizing geospatial social media data to calculate similarities between users, according to their historical venue check-ins. We aimed at predicting the preference levels for all venues that an user has never been to, by comparing their historical check-ins with those users similar to them and those other check-in users they follow. In addition, we intended to identify groups of similar users through spectral clustering, which is widely applied for high-dimensional data. Please refer to our GitHub repo <https://github.com/fanshi118/Time-Out-New-York-MLC> for details of our work.

Data

Since the Foursquare API does not provide public access to individual check-ins, we need to get such data indirectly from those Foursquare check-in records shared on Twitter. The data that we acquired were a sample of such records from Feb 2014 to Feb 2015, parsed using PySpark from 113 GB geotagged tweets around New York City (refer to Acknowledgement for information on data source). We were able to get features including user ID, name, and bio, as well as check-in message, time, and location (latitude, longitude, and city name).

We used the Foursquare Venue Search API to scrape venue information including location ID, name, and category for each unique location denoted by latitude and longitude. Having removed records outside of NYC and venues irrelevant to recreational activities (i.e.

airports, banks, bus stops), we kept all users with at least 5 check-ins throughout the 13 months. The cleaned dataset consists of 4,365 users, 838 venues, and 54,017 check-ins.

Methodology

In a traditional user-item recommendation system, such approach that gives personalized recommendations is called Collaborative Filtering (CF). Many online web services such as Yelp and Yellowpage allow users to express preferences for locations using ratings and then share location preferences with similar users. According to an article from Microsoft Research [3], three types of operations are undertaken in most of the CF-based location recommender systems: similarity inference, candidate selection, and recommendation score prediction. To be more specific, the system should 1) calculate a similarity score between users using their historical ratings on locations; 2) select candidate locations using the user's current venues; and 3) predict the rating that a user may give to a location. A user's location history can be categorized as online rating history of locations and check-in history in location-based social networks. In our project, we want to predict how likely a user may go to a place he has never been using his historical check-in record.

In this section we will introduce the techniques we used in memory-based collaborative filtering and trust-based venue recommendation [4]. Then we use spectral clustering to identify the group similarities, in addition to individual ones.

- Memory-Based Collaborative Filtering

Let U and V denote the set of targeted users and the venues in our scenario. The check-in activity a user $u_i \in U$ has at a venue $v_j \in V$ is described as $c_{i,j}$. $c_{i,j} = 1$ represents user u_i has a check-in at venue v_j before, and $c_{i,j} = 0$ means there is no record of user u_i visiting venue v_j before. The record $c_{i,j}$ are therefore used to discover the user's preference toward a place, which can be used to predict how likely the user would like to go to an unvisited venue. We define this probability by $\hat{c}_{i,j}$ and obtain this check-in probability of user u_i visiting v_j as follows:

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

,

where $w_{i,k}$ is the similarity weight between users u_i and u_j.

To compute the similarity weights between users u_i and u_j, we choose cosine similarity weight between users u_i and u_j [5], denoted as $w_{i,j}$, which is defined as follows:

$$Similarity = \frac{\sum_{K \in V} c_{i,k} \times c_{j,k}}{\sqrt{\sum_{k \in V} c_{i,k}^2} \times \sqrt{\sum_{k \in V} c_{j,k}^2}}$$

- Trust-Based Venue Recommendation

From a social media user's standpoint, following certain users may indicate a tendency of referring to their online activities. We suppose such patterns exist when it comes to venue exploration. Therefore, 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. Essentially, we established a trust-based network in which the relationships between users are denoted by 0s and 1s.

Let $A : a(i,j)$ denote the user relationship matrix. The relationship $a(i,j)$ between two user u_i and u_j equals 1 if user u_i follows user u_j on twitter and $a(i,j) = 0$ can represent as no following relationship from user u_i toward user u_j. We define the trust-based similarity weight $\tilde{w}_{i,j}$ as following:

$$Combined\ Similarity = Similarity \times A$$

We also define trust-based venue recommendation probability as follows:

$$\tilde{c}_{i,j} = \sum_{u_k \in V} (w_{i,k} + \hat{w}_{i,k}) \times c_{k,j}$$

,

where $w_{i,k}$ is the check-in similarity weight and $\tilde{w}_{i,k}$ is the relationship similarity weight.

- Spectral Clustering

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, which is defined as follows :

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

,

where c_j is the j[th] cluster of users, namely,

$$c_i = \left\{ u_{i_1}, u_{i_2}, \dots, u_{i_{k_i}} \right\}$$

We framed our analysis based on the three techniques described above and divided the tasks in such a way that

- Problem identification and literature review—Tianyi, Xiaoge
- Data collection and extraction—Shi, Yuxiang
- Data analysis—Yuxiang, Tianyi, Xiaoge, Shi
- Summary of results and findings—Yuxiang, Tianyi, Xiaoge, Shi

Outcomes

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. Instead of showing all the ranked lists, we picked two user cases randomly to specify our findings and make comparisons between two recommendation approaches.

Table 1 shows one example of the user historical check-in record. The user was randomly chosen from 4,365 users from our dataset, who showed a fond of different types of food and public places like basketball stadium and history museum. With the memory-based collaborative filtering, the suggested places (Table 2) that the user has not checked-in before include four public places (Barclays Center, Yankee Stadium, Central Park, the Met, and Columbus Circle) and one store (Disney store). These places appear on the list for that they have the most user check-in records which makes a collaboratively significant priority on the user's list of potential venues. We expect the trust-based recommendation could improve the recommendation result and target the should-go places to be more personalized. But actually we can only find one person that he is following on Twitter. Table 3 shows the venue list from the trust-based algorithm, which is essentially the same list as Table 2, but with relatively lower predicted check-in probability as well as lower standard deviation.

The main observation from comparing Table 2 and 3 is that trust-based CF did not change the recommendation result undesirably but decrease the standard deviation, which could be counted as a subtle improvement in the confidence level of predicted check-in probability.

Place Visited	Category	Check-in Times
Thai Son	Vietnamese Restaurant	1
Junior's Restaurant & Bakery	American Restaurant	1
Madison Square Garden	Basketball Stadium	1
Intrepid Sea, Air & Space Museum	History Museum	1
Ellen's Stardust Diner	Diner	2
Bubba Gump Shrimp Co.	Cajun / Creole Restaurant	1

Table 1: User Historical Check-in Record Sample

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

Table 2: 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

Table 3: Trust-Based Venue Recommendation Sample

In another case which was showed in Table 4, the user has checked-in more venues from the same category, like bars and lounges, and received a venue list which as well consists of bars and lounges. Compared with the first case, we can find out that the effectiveness of the trust-based recommendation system behaves differently with respect to user's different check-in history and the number of related users.

Our clustering analysis discover a monotonous decreasing trend between the number of clusters and the average in-cluster similarity, as indicated by Figure 1. With more cluster centers included, each cluster tends to be more equally distributed with an attribute of less extremely high similarity and a less overall average similarity.

Place Visited	Category	Place Recommended	Category
Studio 450	Event Space	The 13th Step	Harlem Tavern
Mercury Bar	Bar	Delicatessen	American Restaurant
Sinigual Contemporary Mexican Cuisine	Mexican Restaurant	Cafeteria	New American Restaurant
Oasis	Falafel Restaurant	Harlem Tavern	Music Venue
Cascabel Taqueria	Mexican Restaurant	Pacha NYC	Nightclub
The DL	Lounge	Mother's Ruin	Cocktail Bar
Sweet and Vicious	Lounge	Radegast Hall & Biergarten	Beer Garden

Table 4: Another Example of Trust-Based Recommendation

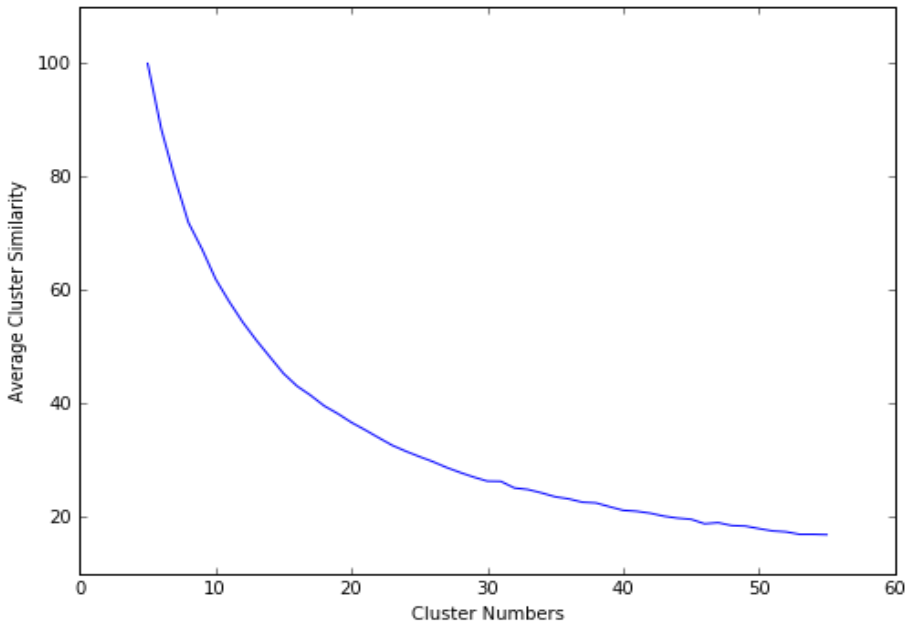

Figure 1. Average Cluster Similarity vs Number of Clusters

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. We believe that such features can boost the overall robustness of our approach for the following reasons: check-in time—temporal dynamics can be an important component in capturing real-time user activity patterns; venue category—instead of

looking venue by venue, it may be tempting to group the venues, so the check-in records are contextually captured; user bio—we were barely able to get personal information on each user (i.e. age, gender, profession), so some text mining on user bios may be an alternative approach.

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. It would have been worthwhile to implement it, have we had enough time to delve deeper into the analysis.

Acknowledgements

The Twitter data were collected for research and academic purposes from the Twitter's Public API by the Visualization and Data Analysis lab at New York University (VIDA-NYU). Special thanks to Professor Huy T. Vo for making the data available to us.

References

1. Cranshaw, Justin, et al. "The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City." *Sixth International AAAI Conference*, pp. 58-65. Accessed on 30 April, 2016. Web. URL: <https://s3.amazonaws.com/livehoods/livehoods_icwsm12.pdf>
2. Ricci, Francesco, et al. "Introduction to Recommender Systems Handbook." *Recommender Systems Handbook*, pp. 1-35. Accessed on 5 October, 2010. Web. URL: <http://link.springer.com/chapter/10.1007%2F978-0-387-85820-3_1#page-1>
3. Bao, Jie, et al. "Recommendations in location-based social networks: a survey." *Geoinformatica*, 6 February 2015, pp. 525-565. Accessed on 30 April, 2016. Web. URL: <http://research.microsoft.com/pubs/191797/LBSN-survey.pdf>
4. Pham, Manh Cuong, et al. "A Clustering Approach for Collaborative Filtering Recommendation Using Social Network Analysis." *Journal of Universal Computer Science*, vol. 17, no. 4, pp. 583-604. Accessed on 30 April, 2016. Web. URL: <http://www.jucs.org/jucs_17_4/a_clustering_approach_for/jucs_17_04_0583_0604_pham.pdf>
5. 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. Accessed on 30 April, 2016. Web. URL: <http://pdfs.semanticscholar.org/e06d/c0b64137a555ed68f40071477e75a87b0e12.pdf>
6. Luxburg, Ulrike von. "A Tutorial on Spectral Clustering." *Springer*, 17 (4), 2007. Accessed on 30 April, 2016. Web. URL: <http://www.kyb.mpg.de/fileadmin/user_upload/files/publications/attachments/Luxburg07_tutorial_4488%5b0%5d.pdf>