We Know Your Preferences in New Cities: Mining and Modeling the Behavior of Travelers

Rong Xie, Yang Chen, Qinge Xie, Yu Xiao, and Xin Wang

The authors investigate travelers' preferences based on the check-in data collected from a popular location-based social application called Swarm. They conduct a thorough analysis of the check-in data to discover the variation in travelers' preferences between cities with different characteristics, and to build a model for predicting the venue types of travelers' interests in each city.

Abstract

The trend of globalization motivates people to travel more often to different cities. In order to provide better suggestions for travelers, it is important to understand their preferences for venue types. In this article, we investigate travelers' preferences based on the check-in data collected from a popular location-based social application called Swarm. We conduct a thorough analysis of the check-in data to discover the variation in travelers' preferences between cities with different characteristics, and to build a model for predicting the venue types of travelers' interests in each city. Our experimental results demonstrate that the F1-score increases by 0.19 when taking into account the characteristics of the destination city. Moreover, our approach outperforms collaborative filtering, a widely used approach to the design of recommendation systems.

Introduction

With the rapid development and increasing popularity of location-based social applications (LBSAs), such as Foursquare's Swarm [1, 2], Skout [3], and Momo [4], more and more check-in data become available for analyzing people's mobility behavior in urban areas. A check-in contains information about a visit, including the time and the visited venue. A massive set of check-in data can reveal users' preferences for venue types, which facilitates more accurate recommendations.

In this work, we aim to find out the main influencing factors of travelers'[1] preferences for venue types based on a massive set of check-in data collected from Foursquare's Swarm app, a leading LBSA. The information of a home city is obtained from Swarm users' profiles if available, or inferred from their check-in history. We call the city to be visited the *destination city*. Our dataset contains more than 33 million check-ins of nearly 20,000 Swarm users across the world. We achieve our goal in the following two steps.

First, we investigate the factors that affect travelers' preferences for venue types in the destination city by investigating the questions: Do travelers and local people have similar preferences? Do a traveler's preferences for venue types vary between different cities? When a person is traveling, does she still have the same preferences as in her home city?

Second, based on the findings of the first step, we build a model that quantifies the impact of different factors, and evaluate the model with a real-world dataset collected from Swarm.

Overall, our key findings and contributions can be summarized as below.

•By performing a thorough analysis of check-ins generated in New York City (NYC), San Francisco (SF), and Hong Kong (HK), we find out that a traveler's preferences for venue types are more similar to those of other travelers rather than local people. Moreover, travelers adapt their preferences for venue types to the characteristics of the destination city. Regarding the city characteristics, San Francisco, for example, is featured as a seaport city, full of delicious seafood; it is also the home of many information technology companies, such as Google and Facebook. To the best of our knowledge, this is the first work that considers the characteristics of the destination city in the prediction of travelers' preferences.

•Our model describes the influencing factors of the variation in travelers' preferences between different cities. The model can be used for predicting travelers' preferences for venue types. Our experimental results demonstrate that the F1-score increases by 0.19 by taking the characteristics of the destination city into account. Moreover, our approach outperforms collaborative filtering, a widely used approach to the design of recommendation systems.

We believe our work can facilitate a deeper understanding of travelers' behavior in destination cities. Moreover, it can be used by LBSA service providers to offer better recommendation services.

The rest of this article is structured as follows. We first review the related work, and introduce the Swarm dataset. Next, we explain how people's preferences for venue types are related to their roles (i.e., local people or travelers) and the city characteristics. Finally, we build a model for predicting travelers' preferences before we conclude the whole work.

Related Work

Human mobility behavior using social media data has been widely studied. Zhang *et al.* [5] investigated urban dynamics by correlating the location,

[1] We define a *traveler* as a person who has stayed consecutively for 1 to 15 days in a city which is not her home city.

Digital Object Identifier:
10.1109/MCOM.2018.1700581

Rong Xie, Yang Chen, Qinge Xie, and Xin Wang are with the School of Computer Science, Fudan University, and the Shanghai Key Lab of Intelligent Information Processing, Fudan University, and the State Key Laboratory of Integrated Services Networks, Xidian University; Yu Xiao is with Aalto University.

time, and activity information extracted from geotagged tweets and Foursquare check-ins. Cranshaw *et al.* [6] studied the dynamics, structure, and characteristics of cities by mining Foursquare check-ins using a clustering method. Additionally, Cho *et al.* [7] studied human mobility patterns based on both cell phone location data and the check-ins collected from two LBSAs, Gowalla and Brightkite. They found that humans made periodic movements that were geographically limited as well as random jumps correlated with their social networks. Hummel and Hess [8] identified human movement activities, such as shopping and traveling, by studying the GPS traces. Atzmueller *et al.* [9] studied users' visiting behavior toward performances in an event by using real-world data collected at a music event in Munich. By using graph analysis, they found that users and performances had a strong correlation, and users had strong individuality as well. In this work, we focus on travelers' preferences for venue types, taking into account both human mobility and the characteristics of the destination cities.

Some previous works tried to make venue recommendations for travelers. Bao *et al.* [10] designed and implemented a recommendation system, taking into account the user's preferences for venue types, opinions of local expertise, and the user's current location. They first identified experts who share similar interests of venue types with the user, and then applied collaborative filtering to predict the user's final ratings for unvisited venues. When selecting experts, they did not distinguish local people from travelers. Meanwhile, according to our study, a traveler's preferences are actually more similar to those of other travelers than those of local people. Pham *et al.* [11] assumed that people often visit a set of venues close to each other instead of a single venue. Accordingly, they proposed to recommend a region having multiple attractive points of interest (POIs) to travelers. Ference *et al.* [12], on the other hand, considered friends' preferences in addition to the user's preferences and geographical proximity. They combined the above three factors into a CF-based model for recommending venues near the user's current location. Moreover, Çelikten *et al.* [13] proposed to match similar regions across different cities and to provide out-of-town recommendations based on users' preferences in their home cities. Similar to previous works, we utilize the check-ins collected from a popular LBSA for analyzing the influencing factors of travelers' preferences. Differently, we take the city characteristics into account when making recommendations. What is more, our focus is placed on preferences for venue types rather than exact venues. To the best of our knowledge, there is no literature that models travelers' preferences for venue types in different destination cities.

Data Collection

As a representative LBSA, Swarm users generate more than 8 million check-ins each day,[2] whereas each user's check-ins are only visible to the user's Swarm friends. From April 15, 2017 to May 20, 2017, we used 10 Swarm accounts owned by our team members to send friend requests to randomly selected Swarm users. Each of the requests includes the following message: "We are from Fudan University in China. We add friends in order to do our research on user behavior modeling. Your data will be used for research purposes only, and will never be shared with a third party." Eventually, 19,484 Swarm users agreed to join our experiment, and we were able to crawl the complete check-in history of these users.

We collected in total 33,178,113 check-ins made in 2616 cities in 184 countries. Among these cities, we selected the check-ins generated in NYC, SF, and HK for detailed analysis. Due to space limitations, we had to choose a few cities to present in the article. We selected NYC, SF and HK based on the following intuition. First, the number of check-ins generated by travelers in any of these selected cities ranks among the top 10 in the world. Second, the selected cities are all popular metropolises but with different characteristics: NYC is an international metropolis with a large number of urban population; SF is the cultural, commercial, and financial center of northern California and a popular tourist destination; HK is a well-known Asian city with a different cultural background than NYC and SF. The similarities and differences between these three cities facilitate a more complete impact analysis of travelers' preferences for venue types. For future work, we plan to take more cities for comparison. Besides check-ins, our dataset also contains the users' demographic information, including gender, home city, and number of friends.

The similarities and differences between these three cities facilitate a more complete impact analysis of travelers' preferences for venue types. For future work, we plan to take more cities for comparison. Besides check-ins, our dataset also contains the users' demographic information, including gender, home city and the number of friends.

Analysis of Travelers' Preferences

Previous works [10, 12, 13] usually make recommendations for people based on their visiting history. For example, they assume that people who have shared similar preferences for venue types in the past would prefer to visit the same types of venues in the future, no matter whether they are local people or travelers in the same city. To figure out whether the assumption holds, we conduct a thorough analysis of Swarm check-ins, trying to find answers to the following questions:

- Do travelers and local people in the same city share similar preferences?
- Do a traveler's preferences for venue types vary between different cities?
- When a person is traveling in other cities, does she still have the same preferences as in her home city?

Spatiotemporal Check-in Distributions

To answer the first question, we compare preferences between travelers and local people in NYC, SF, and HK. In the case of NYC, our dataset includes 593,557 check-ins of local people and 149,112 check-ins of travelers. The numbers of check-ins are 167,666 for local people and 26,117 for travelers in SF, and 335,720 for local people and 38,631 for travelers in HK. The travelers in each city come from all over the world. We investigate the spatiotemporal mobility patterns of Swarm users based on these check-ins.

Figure 1 describes the distribution of check-ins at different times of day in NYC, SF, and HK, respectively. We divide a day into four even periods, which are 12 a.m.–6 a.m., 6 a.m.–12 p.m., 12 p.m.–6 p.m., and 6 p.m.–12 a.m. We compare the percentages of check-ins made in each time period between travelers and local people,

[2] https://venturebeat.com/2016/03/24/foursquares-swarm-now-lets-you-make-sense-of-where-youve-been/, accessed October 10, 2017.

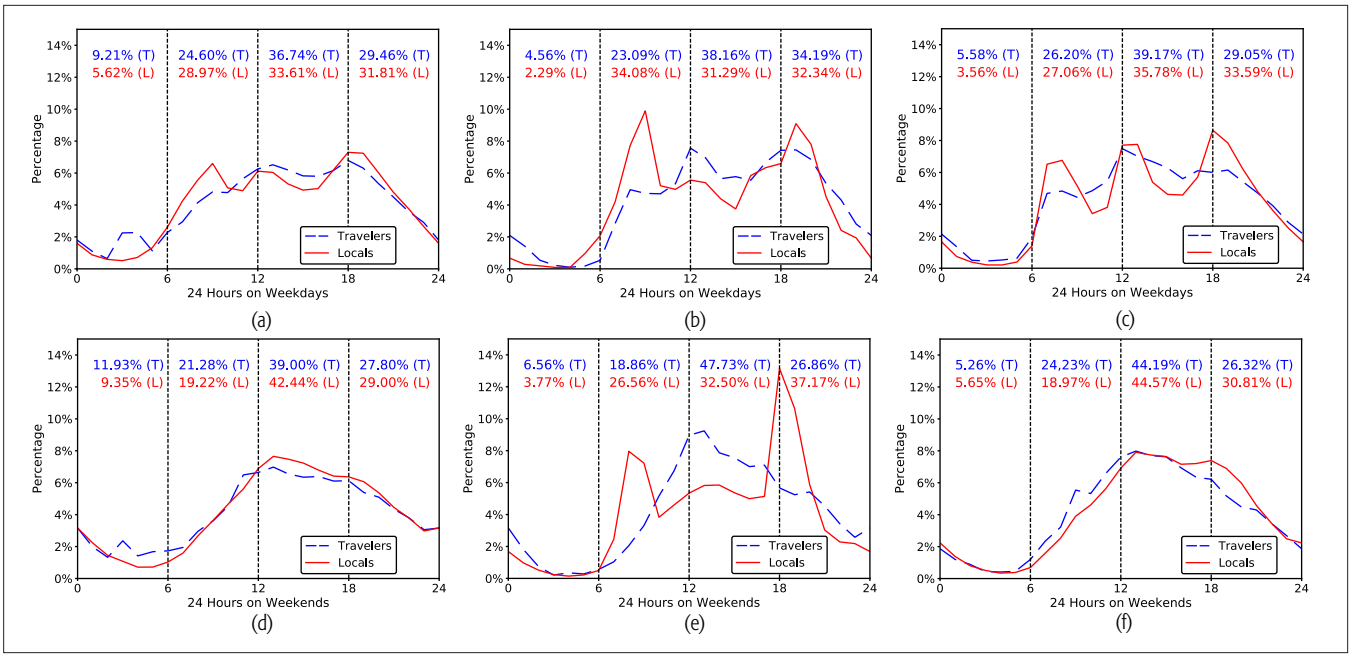

Figure 1. Temporal distribution of check-ins in NYC, SF, and HK during a week: a) NYC-weekdays; b) SF-weekdays; c) HK-weekdays; d) NYC-weekends; e) SF-weekends; f) HK-weekends.

and between weekdays and weekends. We find that both travelers and local people are in general inactive between 12 a.m. and 6 a.m., and very active from 1 p.m. to 9 p.m.. Looking into the details of the temporal distribution, the differences in the check-in activities between travelers and local people are summarized below.

First, the number of check-ins made by travelers in NYC reaches a peak at around 3 a.m. on weekdays, when local people are the least active in generating check-ins. Obviously, travelers are more involved in the night life activities than local people during weekdays.

Second, on weekdays, in all three cities, the check-in percentages between 6 a.m. and 12 p.m. of local people are all higher than those of travelers. The check-in percentage differences in this time period between local people and travelers in NYC, SF, and HK are 4.37, 10.99, and 0.86 percent, respectively. This is because local people go to work in the morning and start publishing check-ins for example, on the way to work.

Third, on weekdays, in all three cities, travelers' check-ins distribute more evenly than local people's between 9 a.m. and 7 p.m. Comparing the maximum percentage of check-ins in each hour with the minimum one, the average value of the differences of three cities in the case of travelers is 3.95 percent, lower than that of local people, which is 5.22 percent. This is because local people are relatively more active during the break at noon and after work (around 6 p.m.). Finally, although the temporal pattern of check-in activity varies between weekdays and weekends for both travelers and local people, the difference between travelers and local people during weekends is small since they are all experiencing leisure time.

We visualize the spatial distribution of check-ins in the form of heat maps in Fig. 2. Darker colors represent higher percentages of check-ins made in the corresponding areas. We can see the difference between travelers and local people. For example, travelers' check-ins in NYC are highly concentrated in midtown, lower Manhattan, and Utica Avenue in Brooklyn, while local people's check-ins are more distributed across areas like Manhattan, Long Island City, and Forest Hills. In SF, travelers make the most check-ins at tourist attractions (the northeastern part of SF), such as Alamo Square, Union Square, and Fisherman's Wharf. However, local people prefer to make check-ins at supermarkets, apartments, coffee shops, and nightclubs in their daily lives. In HK, travelers prefer to check in at large shopping malls, such as Elements, as well as tourist attractions, such as Mongkok and Victoria Peak. In contrast, local people there often perform check-ins at bus stations. Overall, travelers prefer to perform check-ins at tourist attractions such as Victoria Peak in HK and venues having distinctive features such as Manhattan in NYC, whereas local people have more check-ins in places they often visit in their daily lives, such as supermarkets.

To highlight the difference in the spatial distribution, we calculate the Spearman's rank correlation coefficient of two 20-dimensional vectors, V_{local} and $V_{traveler}$, for each city. Each dimension in a vector refers to the percentage of check-ins made by local people or travelers in the corresponding type of venues. The correlation coefficient measures the similarity of the popularity rank of the venue types between local people and travelers. Its value is between –1 and 1, and the closer to 0, the less similar. The correlation coefficients between local people and travelers in NYC, SF, and HK are 0.36, 0.24, and 0.17, respectively. The small values of correlation coefficients clearly reflect the significant differences in preferences for venue types between local people and travelers.

In summary, check-ins of travelers and local people follow different temporal and spatial distributions. Regarding the venue types, travelers prefer tourist attractions, shopping malls, and other

venues having distinctive features, whereas local people would be more likely to make check-ins at places they visit regularly, such as supermarkets, coffee shops, and bus stops.

Variation in Travelers' Preferences

To figure out whether a traveler's preferences vary between different cities, we compare travelers' preferences for venue types in NYC, SF, and HK. Table 1 lists the 20 most popular venue types according to the percentage of check-ins made in each venue type by travelers in each of the three cities.[3]

We find out that the characteristics of the city impact the ranking of venue types there. For example, "Shopping Malls" ranks as the second most popular venue type in HK, whereas it does not appear in the list of the other two cities. It seems that people prefer to travel to HK for shopping, as HK is famous for being "the paradise of shopping." Similarly, "Banks" is included in the list of NYC, due to NYC's role as a financial center. Meanwhile, "Roads" and "Parks" rank high in the list of NYC because many important commercial streets there belongs to "Roads," such as Fifth Avenue and Utica Avenue. This implies that travelers prefer to go to such commercial streets for sightseeing or shopping. Also, Central Park is a popular attraction in NYC. In the case of SF, there are more check-ins at "Historic Sites" than in HK and NYC. Transamerica Pyramid, Fisherman's Wharf, and Coit Tower are all famous historic sites in SF. The high rank of "Piers" in HK and SF reflects the characteristics of seaport cities. Besides, preferred restaurants reflect the demographic profiling of the cities. Specifically, "Asian Restaurants" ranks high in all three cities. HK is an Asian city, so it is not a surprise that Asian food is popular there. For NYC and SF, the large Asian populations drive the popularity of Asian food. In addition, "American Restaurants" ranks high in both of the American cities. "Seafood Restaurant" is one of the most popular restaurant types in SF, showing the advantage of its location.

Based on our observation, we can conclude that the characteristics of a city impact travelers' preferences. In other words, travelers adapt their preferences to the characteristics of the destination city. Therefore, it is important to take the city characteristics into account when trying to understand travelers' preferences for venue types. Also, the city characteristics could be reflected by travelers' preferences there.

The Individual User's Cross-City Preferences

In this subsection, we answer the third question, which is whether an individual changes her preferences when she travels to other cities.[4]

To compare a user's preferences in the home city h with that in another chosen city o, we define a *cross-city preference resemblance metric*, *Resemblance* (u_h, u_o). It is a variant of the user resemblance calculation approach proposed in [10]. The parameters u_h and u_o represent the user u's preferences in the home city h and the chosen city o, respectively. We extract the resemblance from the venue type information in the user's visit history, and calculate *Resemblance* (u_h, u_o) based on the following two metrics:

1. *User preference similarity Similarity* (u_h, u_o)

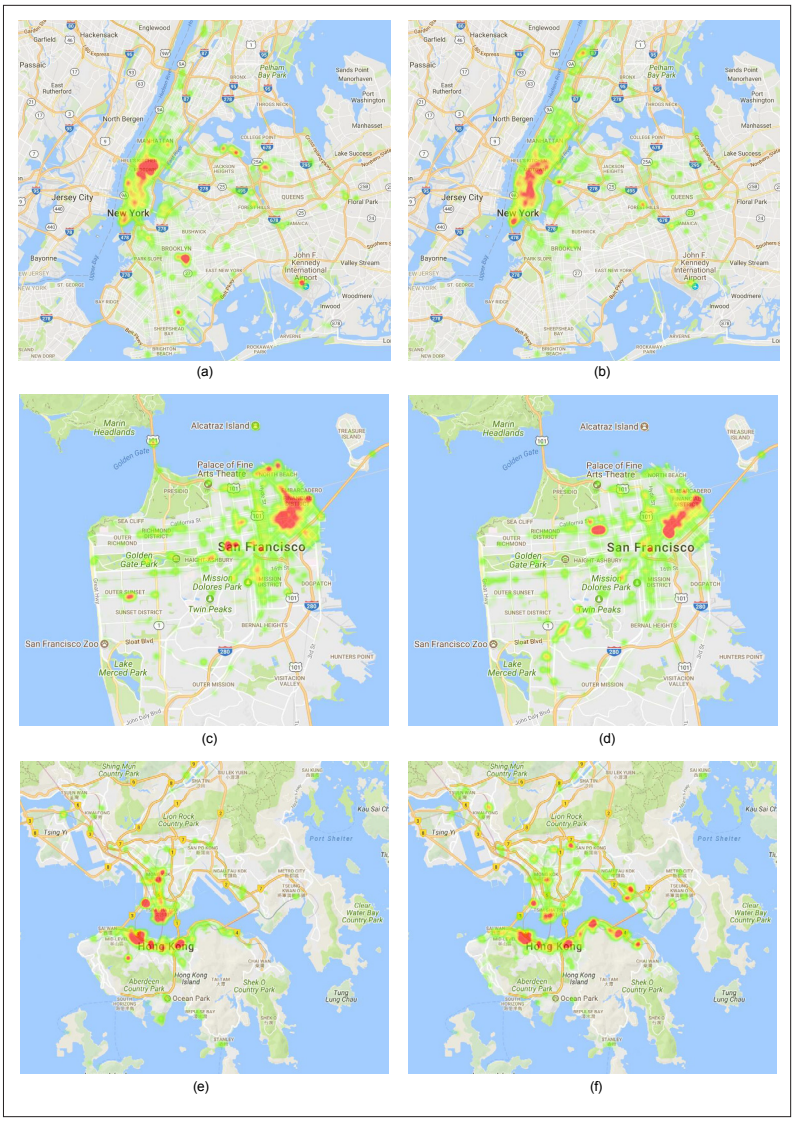

Figure 2. Spatial distribution of check-ins in NYC, SF, and HK: a) NYC-travelers; b) NYC-locals; c) SF-travelers; d) SF-locals; e) HK-travelers; f) HK-locals.

2. *User preference entropy Entropy* (u_c)

Here u_c represents the user u's preferences in city c. *Resemblance*(u_h, u_o) is defined as

$$Resemblance(u_h,u_o)=\frac{Similarity(u_h,u_o)}{1+\left|Entropy(u_h)-Entropy(u_o)\right|}.$$

The value of *Resemblance*(u_h, u_o) would be high if the user shows similar preferences in these two cities, and the difference in the preference diversity (entropy) is small between these cities.

Calculation of Similarity (u_h, u_o): First, we calculate a group of term frequency-inverse document frequency (TF-IDF) values [14] for city h and o, respectively. Each element in one group is the TF-IDF value of a venue type that the user has visited in the city. We denote a user u's check-in percentage of a venue type t as $TF(u, t)$ and denote the natural logarithm of a value about t as $IDF(t)$. That value is the total number of people in the city divided by the number of people who have visited venue type t in that city. Then user u's TF-IDF value for venue type t in this city would be $TF(u, t) \times IDF(t)$. This value represents

[3] Swarm defines five levels of venue types from the most coarse-grained one to the most fine-grained one. For instance, "Food" is a level one category, while "American Restaurants" is a level two category belonging to the "Food" category. The venue type recorded by each check-in could belong to any of the five levels. To standardize the level of check-in venues, we first calculate the distribution of all check-ins among the five levels and find out that 74 percent of them use level two venue types. Therefore, we merge all venues from level three to level five into level two and discard the check-ins at level one venues (covering 0.2 percent of the whole set).

[4] The same as in the previous subsection, we discard the check-ins that provide only level one venue types and merge venue types of all the other levels into level two.

An individual's preferences for venue types in her home city are different from those in the destination city. Therefore, we should take the destination city's characteristics into consideration when making recommendations of venue types for travelers. Also, we should differentiate local people from travelers when trying to find similar users.

Rank	HK	Percentages	NYC	Percentages	SF	Percentages
1	Asian restaurants	24.91%	Bars	11.91%	Bars	18.73%
2	Shopping malls	21.41%	Roads	9.56%	Asian restaurants	13.79%
3	Housing developments	6.68%	Parks	8.17%	Offices	9.95%
4	Bars	6.38%	Asian restaurants	7.55%	Parks	7.49%
5	Piers	4.94%	Offices	7.50%	Food and drink shops	6.35%
6	Theme parks	4.79%	Food and drink shops	7.31%	Mexican restaurants	4.59%
7	Clothing stores	4.02%	Athletics and sports	5.15%	American restaurants	4.45%
8	Bakeries	3.50%	Plazas	4.56%	Seafood restaurants	3.38%
9	Food and drink shops	3.30%	American restaurants	4.11%	Clothing stores	3.20%
10	Dessert shops	3.19%	Other great outdoors	4.04%	Athletics and sports	3.14%
11	Fast food restaurants	2.20%	Clothing stores	3.47%	Piers	2.91%
12	Italian restaurants	2.04%	Museums	3.31%	Plazas	2.87%
13	Government buildings	1.91%	Delis/bodegas	3.29%	Dessert shops	2.80%
14	Electronics stores	1.70%	Pharmacies	3.22%	Bakeries	2.71%
15	Parks	1.65%	Italian restaurants	2.97%	Pizza places	2.52%
16	Spiritual centers	1.58%	Bakeries	2.85%	Museums	2.43%
17	Roads	1.56%	Banks	2.83%	Burger joints	2.33%
18	Department stores	1.52%	Bridges	2.77%	Italian restaurants	2.27%
19	Athletics and sports	1.42%	Burger joints	2.76%	Sandwich places	2.04%
20	Offices	1.31%	Pizza places	2.66%	Historic sites	2.03%

Table 1. Top 20 venue types in HK, NYC, and SF.

the degree of a user's prominent preferences for that venue type. Then we calculate *Similarity*(u_h, u_o). For all the venue types in city *h* and *o*, we first select the smaller TF-IDF value from the two TF-IDF values of the same venue type in the two cities. Then we sum up all the selected smaller TF-IDF values to get *Similarity*(u_h, u_o).

Calculation of Entropy (u_c): We denote a user *u*'s check-in percentage of venue type *t* in city *c* as $p(u_c, t)$ and denote the venue type set in city *c* as *t*. Then the *Entropy*(u_c) is calculated as $-\Sigma_{t \in T} p(u_c, t) \log p(u_c, t)$. Entropy is able to capture the diversity of a user's preferences in a city.

Calculation of Resemblance (u_h, u_o): We divide the *Similarity*(u_h, u_o) by the absolute value of the difference in the user preference entropy between city *h* and city *o*. To prevent the denominator from being 0, we add an extra 1 to it as in [10]. Furthermore, to make the value more expressive, we normalize each *Resemblance* (u_h, u_o) by dividing it by the maximal value of *Resemblance* (u_h, u_o).

Figure 3 is the cumulative distribution function (CDF) of the normalized cross-city preference resemblance *Resemblance* (u_h, u_o). The solid line shows the distribution of users' average cross-city preference resemblance. Each value is calculated as the sum of the user's cross-city preference resemblance between home city and one chosen city divided by the number of all the chosen cities. We choose all the cities where the user has stayed consecutively for 1 to 15 days. The dashed line shows the distribution of users' largest cross-city preference resemblance. More than 95 percent of users have an average preference resemblance smaller than 0.2. More than 75 percent of users have the largest preference resemblance less than 0.2. It shows the significant differences between users' preferences in the home city and in the other cities. In other words, people prefer different venue types when they travel to a different city.

In conclusion, travelers and local people show different preferences in the same city, and the characteristics of each city impact travelers' preferences there. Moreover, an individual's preferences for venue types in her home city are different from that in the destination city. Therefore, we should take the destination city's characteristics into consideration when making recommendations of venue types for travelers. Also, we should differentiate local people from travelers when trying to find similar users.

City Characteristics and Travelers' Preferences

To provide more accurate predictions of travelers' preferences for venue types in different cities, we build a model that quantifies the impact of different factors on a traveler's preferences for venue types. The model can be used for predicting whether the traveler would visit certain venue types (e.g., the 10 most popular venue types chosen by travelers) in the destination city. The variables of the model are defined based on three categories of features.

First, demographic features of the traveler,

including gender, home city, and home country. The feature value is set to none if the traveler does not provide the information. We encode each feature value using one-hot encoding, where one-hot is a group of bits with only one single high (1) bit and all the others low (0).

Second, features of user-generated contents, including the number of the traveler's friends in the Swarm network, the total number of check-ins made by the traveler, and the traveler's TF-IDF values of each visited venue type in her home city. The TF-IDF values represent the traveler's prominent preferences for venue types in her home city.

Third, the features that represent the characteristics of the destination city from travelers' perspectives. As travelers may have different opinions about a city's characteristics, depending on their personal interests, we propose to apply *K*-means clustering to classify travelers in the same destination city by their preferences for venue types into *K* groups, and to summarize the characteristics of the destination city from each group's perspective. The travelers within the same group are expected to have common interests. The features we use are the average visit frequencies (excluding the check-in data of the traveler being predicted) of each group for each venue type in the destination city.

The value of *K* for each city is selected by the following intuition. First, the value of *K* should be smaller than 4 percent of the number of travelers in the city. The numbers of travelers in NYC, SF, and HK are 1348, 510, and 630, respectively. Second, the value of *K* affects the results of city characteristic analysis, and further the training of the model for preference prediction. Thus, we select the value of *K* that corresponds to the model with the best prediction performance.

For each city, we randomly select 80 percent of travelers and their data for training, using the rest for testing, and train a model for each city, respectively. In detail, we choose the 10 most popular venue types among travelers in the destination city, and describe the travelers' previous visits to these venue types in the city with a 10-element tuple. Each element indicates whether the traveler has published any check-in for the corresponding venue type. The modeling is then transformed into a multi-label classification problem, with each predicted venue type as a label and the prediction of each label as a binary classification problem. We train the classifier using five classic machine learning algorithms, including XGBoost, Random Forest, Decision Tree, Naive Bayes, and support vector machine (SVM). In particular, XGBoost is an emerging scalable tree boosting system that has been widely used in machine learning contests. We measure the prediction performance of each label using precision, recall, and F1-score. Precision is the fraction of users predicted to have been to the venue type who really have been there. Recall is the fraction of users having been to that venue type that are correctly predicted. F1-score is the harmonic mean of precision and recall. The results shown in Table 2 and Fig. 4 are the average over 10 labels.

We compare the prediction performance of the model with and without using the characteristics of the destination city in Table 2. Among

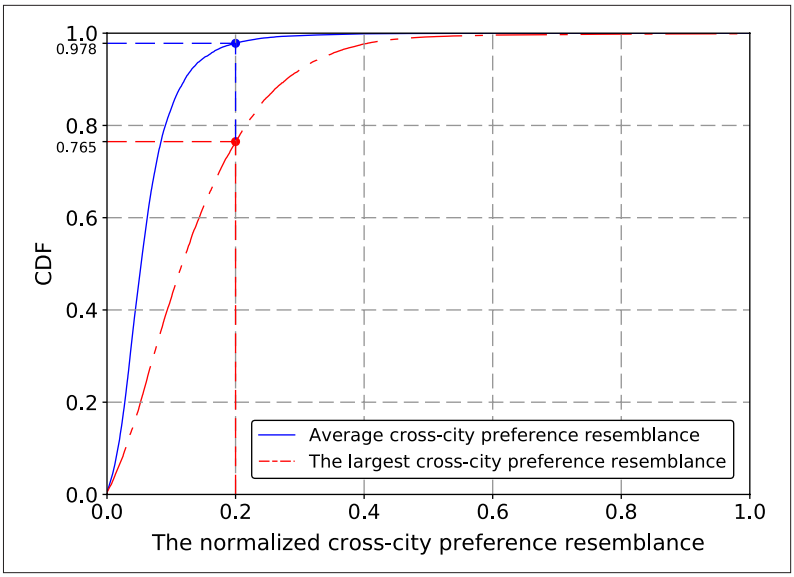

Figure 3. Distribution of normalized users' cross-city preference resemblance.

City	K	Algorithm	Parameter	Recall	Precision	F1-score	Gain
NYC	19	XGBoost	150 trees, max_depth = 5	0.81/0.76	0.67/0.56	0.73/0.64	0.09
		Random Forest	150 trees, 60 features/tree, max/depth = 12	0.72/0.58	0.70/0.61	0.70/0.59	0.11
		Decision Tree (CART)	criterion = gini, max/depth = 10	0.68/0.56	0.64/0.55	0.66/0.55	0.11
		Naive Bayes	–	0.69/0.58	0.59/0.54	0.63/0.55	0.08
		SVM	kernel = rbf, cost parameter C = 8, γ = 0.0003	0.56/0.51	0.55/0.51	0.55/0.51	0.04
SF	15	XGBoost	150 trees, max/depth = 5	0.82/0.73	0.67/0.51	0.73/0.59	0.14
		Random Forest	150 trees, 59 features/tree, max/depth = 10	0.64/0.43	0.69/0.57	0.66/0.47	0.19
		Decision Tree (CART)	criterion = gini, max/depth = 6	0.69/0.57	0.63/0.48	0.65/0.51	0.14
		Naive Bayes	–	0.71/0.58	0.52/0.44	0.58/0.48	0.10
		SVM	kernel = rbf, cost parameter C = 2, γ = 0.0003	0.54/0.44	0.53/0.45	0.52/0.44	0.08
HK	13	XGBoost	150 trees, max/depth = 4	0.79/0.71	0.74/0.60	0.76/0.64	0.12
		Random Forest	150 trees, 59 features/tree, max/depth = 10	0.72/0.54	0.77/0.66	0.74/0.56	0.18
		Decision Tree (CART)	criterion = gini, max/depth = 7	0.75/0.59	0.66/0.57	0.70/0.57	0.13
		Naive Bayes	–	0.67/0.56	0.61/0.55	0.62/0.53	0.09
		SVM	kernel = rbf, cost parameter C = 2, γ = 0.0003	0.56/0.50	0.60/0.54	0.58/0.51	0.07

Table 2. Comparison of prediction performance. The table compares the precision, recall, and F1-score of the models with and without the feature of city characteristics. The column "Gain" shows the F1-score's improvement after using this feature.

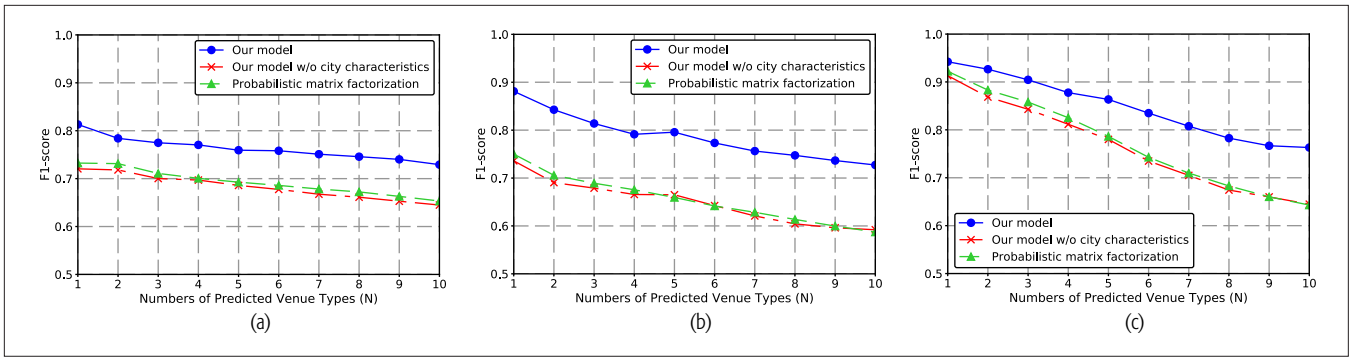

Figure 4. Prediction results of top *N* venue types in NYC, SF, and HK: a) NYC F1-score; b) SF F1-score; c) HK F1-score.

all the machine learning algorithms, XGBoost achieves the best performance. The highest F1-score is 0.76, achieved in the case of HK. The F1-score increases in all three cases by taking into account the city characteristics. The highest gain is obtained in the case of SF, which is 0.19. In general, we observe higher gains in SF and HK than in NYC. These findings are consistent with the results shown in Fig. 4, which shows the average F1-score of the *N* most popular venue types among travelers in the city.

We also compare our approach with collaborative filtering (CF). CF is a method of predicting a user's interests based on the preferences of other users. It has been widely used for location recommendations [10, 12]. We implemented CF using a classic algorithm called Probabilistic Matrix Factorization (PMF) [15]. The input is the "user-venue types" matrix built from travelers' complete visit history, excluding the visit history in the destination city. Element r_{ij} in the matrix represents the number of visits that user *i* has made to any venue belonging to type *j*. The output is another matrix in which element s_{ij} is 1 or 0, representing whether user *i* would go to any venue of type *j* or not in the destination city. Then we calculate the average of precisions, recalls, and F1-scores of the 10 most popular venue types (the 10 labels). The average F1-scores for NYC, SF, and HK are 0.65, 0.59, and 0.64, respectively, as shown in Fig. 4. These F1-scores are much lower than those of our approach taking into account the characteristics of the destination city, but are close to the results without using the characteristics. To sum up, city characteristics affect travelers' preferences for venue types in the destination city, and our approach outperforms CF.

Conclusion and Future Work

In this work, we analyze travelers' preferences for venue types based on the check-ins of Swarm users. We find that people tend to have different preferences when they are traveling outside their home cities, and travelers do not necessarily share the same preferences as local people. Our study also shows that travelers adapt their preferences to the characteristics of their destinations. By taking into account these characteristics in the prediction of travelers' preferences, the F1-score of our model increases by 0.19.

This work deepens the understanding of travelers' preferences for venue types in different cities. Our results can help LBSA service providers develop more accurate recommendations for travelers, and city governors to make their city more traveler-friendly. In the future, we will extend our work to cover more cities, and will try to apply our solution through collaboration with LBSA service providers and city governments.

Acknowledgment

This work is sponsored by the National Natural Science Foundation of China (No. 61602122, No. 71731004), the Natural Science Foundation of Shanghai (No. 16ZR1402200), and the Shanghai Pujiang Program (No. 16PJ1400700). Yang Chen is the corresponding author.

References

[1] S. Lin *et al.*, "Understanding User Activity Patterns of the Swarm App: A Data-Driven Study," *Proc. ACM UbiComp*, 2017, pp. 125–28.
[2] Y. Chen *et al.*, "Measurement and Analysis of the Swarm Social Network with Tens of Millions of Nodes," *IEEE Access*, 2018.
[3] M. Li *et al.*, "All Your Locations Belong to Us: Breaking Mobile Social Networks for Automated User Location Tracking," *Proc. ACM MobiHoc*, 2014, pp. 43–52.
[4] T. Chen, M. A. Kaafar, and R. Boreli, "The Where and When of Finding New Friends: Analysis of a Location-Based Social Discovery Network," *Proc. ICWSM*, 2013, pp. 61–70.
[5] C. Zhang *et al.*, "Regions, Periods, Activities: Uncovering Urban Dynamics Via Cross-Modal Representation Learning," *Proc. WWW*, 2017, pp. 361–70.
[6] J. Cranshaw *et al.*, "The Livehoods Project: Utilizing Social Media to Understand the Dynamics of A City," *Proc. AAAI*, 2012, pp. 58–65.
[7] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and Mobility: User Movement in Location-Based Social Networks," *Proc. ACM SIGKDD*, 2011, pp. 1082–90.
[8] K. A. Hummel and A. Hess, "Estimating Human Movement Activities for Opportunistic Networking: A Study of Movement Features," *Proc. IEEE WoWMoM*, 2011, pp. 1–7.
[9] M. Atzmueller *et al.*, "Social Event Network Analysis: Structure, Preferences, and Reality," *Proc. IEEE/ACM ASONAM*, 2016, pp. 613–20.
[10] J. Bao, Y. Zheng, and M. F. Mokbel, "Location-Based and Preference-Aware Recommendation Using Sparse Geo-Social Networking Data," *Proc. ACM SIGSPATIAL*, 2012, pp. 199–208.
[11] T.-A. N. Pham, X. Li, and G. Cong, "A General Model for Out-of-Town Region Recommendation," *Proc. WWW*, 2017, pp. 401–10.
[12] G. Ference, M. Ye, and W.-C. Lee, "Location Recommendation for Out-of-Town Users in Location-Based Social Networks," *Proc. ACM CIKM*, 2013, pp. 721–26.
[13] E. Çelikten, G. Le Falher, and M. Mathioudakis, "Modeling Urban Behavior by Mining Geotagged Social Data," *IEEE Trans. Big Data*, vol. 3, no. 2, 2017, pp. 220–33.
[14] G. Salton and C. Buckley, "Term-Weighting Approaches in Automatic Text Retrieval," *Info. Processing & Management*, vol. 24, no. 5, 1988, pp. 513–23.
[15] A. Mnih and R. R. Salakhutdinov, "Probabilistic Matrix Factorization," *Proc. NIPS*, 2008, pp. 1257–64.

Biographies

Rong Xie (xieronglucy@fudan.edu.cn) received her B.S. degree

(with honors) in computer science from Fudan University, Shanghai, China in 2016. She is now a graduate student in computer science at Fudan University. Her research interests include location-based social networks and machine learning.

YANG CHEN (chenyang@fudan.edu.cn) is an associate professor within the School of Computer Science at Fudan University. Before that, he was a postdoctoral associate at the Department of Computer Science, Duke University, and a research associate at the Institute of Computer Science, University of Göttingen, Germany. He received his B.S. and Ph.D. degrees from Tsinghua University in 2004 and 2009, respectively. His research interests include online social networks, Internet architecture, and mobile computing.

QINGE XIE (qgxie17@fudan.edu.cn) received her B.S. degree (with honors) in computer science from Zhejiang University of Technology in 2017. She is now a graduate student in computer science at Fudan University. Her research interests include online social networks and data mining.

YU XIAO (yu.xiao@aalto.fi) received her Ph.D. degree (with distinction) in computer science from Aalto University in January 2012. She is currently an assistant professor in the Department of Communications and Networking, Aalto University, where she leads the mobile cloud computing group. Her research interests include edge computing, mobile crowdsensing, and energy-efficient wireless networking. She was the recipient of a Young Researcher Award from the Finnish Foundation for Technology Promotion in 2017.

XIN WANG (xinw@fudan.edu.cn) received his B.S. degree in information theory and his M.S. degree in communication and electronic systems from Xidian University, China, in 1994 and 1997, respectively. He received his Ph.D. degree in computer science from Shizuoka University, Japan, in 2002. He is currently a professor at Fudan University. His research interests include quality of network service, next-generation network architecture, mobile Internet, and network coding.