Exploring Trajectory-Driven Local Geographic Topics in Foursquare

Xuelian Long
University of Pittsburgh
xul10@pitt.edu

Lei Jin
University of Pittsburgh
lej17@pitt.edu

James Joshi
University of Pittsburgh
jjoshi@pitt.edu

ABSTRACT

The location based social networking services (LBSNSs) are becoming very popular today. In LBSNSs, such as Foursquare, users can explore their places of interests around their current locations, check in at these places to share their locations with their friends, etc. These check-ins contain rich information and imply human mobility patterns; thus, they can greatly facilitate mining and analysis of local geographic topics driven by users' trajectories. The local geographic topics indicate the potential and intrinsic relations among the locations in accordance with users' trajectories. These relations are useful for users in both location and friend recommendations. In this paper, we focus on exploring the local geographic topics through check-ins in Pittsburgh area in Foursquare. We use the Latent Dirichlet Allocation (LDA) model to discover the local geographic topics from the check-ins. We also compare the local geographic topics on weekdays with those at weekends. Our results show that LDA works well in finding the related places of interests.

Author Keywords

Location-based Social Networking Services, *Foursquare*, Geographic Topics, Trajectory

ACM Classification Keywords

J.4 Computer Applications: social and behavior science; H.3.5 Online Information Services: web-based services

General Terms

Measurement, Human Factors

INTRODUCTION

Nowadays, Location-based Social Networking Services (LBSNSs) [1] are becoming more and more popular. Because of the rapid developments of the fast 4th generation mobile networks, the powerful interfaces supporting map services, and the smartphones in which GPS modules are embedded, it is very easy for the mobile users to identify their locations and share them in LBSNSs. In a LBSNS, users can explore the places of interests, check in at their current locations, leave tips or comments and add new friends. Therefore, LBSNSs, such as *Foursquare*, *Facebook Places*, etc., have recently attracted a lot of users by using different mechanisms; further, they find ways to motivate them to share their locations in their systems. For example, *Foursquare* had nearly 20 million users in March 2012 [2]. *Facebook Places* reported 200 million monthly active users creating 2 billion actions tagged with locations by April, 2012 [3].

Sometimes, it may be daunting for users to find the right places of interests in LBSNSs, when they go to a new city and/or are not familiar with the city. Thus, LBSNSs, such as *Foursquare*, have launched categories associated with the venues to facilitate the search. For example, if a user is interested in finding a nearby Mexican restaurant, he can choose to explore the venues with "Mexican" as the category in *Foursquare*. But such help is limited because these categories are static and predefined. Moreover, users are interested in other types of categorizations of the venues, i.e. categorizing the venues in accordance with the crowd level, which we call the *geographic topics* in this paper. For example, users may be more interested in knowing which restaurants people usually go to after shopping at a mall; which cafe is more popular around their current locations. To support such user needs, it is important to provide geographic topics in LBSNSs.

Topic models are very common and useful in text classification. In this paper we propose a topic-model based approach for venue classification based on users' trajectories. The key premise is that the venues that appear together in many users' trajectories will probably be taken as geographic topics. Therefore, the venues in the same trajectory-driven geographic topic are potentially and intrinsically related by the human mobility. These geographic topics can be used to (1) understand users' preferences of venues, (2) recommend venues to users based on previous preference, (3) recommend friends and (4) design business strategies.

In this paper, we use the check-in data related to Pittsburgh area collected from *Foursquare* to explore the trajectory-driven local geographic topics. We employ the Latent Dirichlet Allocation (LDA) approach to discover the local geographic topics. Our main contributions are as follows:

- Our proposed topic model based approach investigates the potential and intrinsic relations among the different venues in Pittsburgh area. Our approach can dynamically categorize the venues in *Foursquare* according to the users' trajectories that indicate the crowds' preferences of the venues. The results of the local geographic topics can be used for recommending both locations and friends to users in LBSNSs.

UbiComp '12, Sep 5-Sep 8, 2012, Pittsburgh, USA.
ACM 978-1-4503-1224-0/12/09.

	Check-ins	Venues	Users
All	813,221	16,461	32,113
Weekdays	574,372	16,222	26,224
Weekends	238,849	13,780	22,868

Table 1. Summary of the Data Set

- We consider the temporal differences in discovering geographic topics. Users' check-in patterns on weekdays are quite different from those at weekends. Thus, we apply our model for users' weekday as well as weekend trajectories separately in this paper and identify the differences of the check-in patterns between them.

- Our data set is collected from *Foursquare* directly. Other data sets used in many current research related to *Foursquare* data analysis are typically gathered from *Twitter*, e.g. [17, 12]; hence, these data sets are less complete because: 1) users may not push every check-in in *Foursquare* to their *Twitter* accounts; and, 2) there are about less than 25% *Foursquare* users who connect their *Foursquare* accounts with their *Twitter* accounts [5]. Thus, our analysis is more comprehensive and accurate because our data set is more complete than the *Twitter* based data sets.

In the rest of the paper, we summarize our data set and then describe the geographic topic modeling. We also present our experiments and discuss some interesting findings based on the local geographic topics generated by our experiments. We review the related work and conclude the paper with a discussion of our future work.

DATA SET

We crawled users' check-in data in Pittsburgh area in *Foursquare* from Feb 24 to May 23, 2012. We define Pittsburgh area as a square with sides of around 40 miles and centered at Pittsburgh Downtown. We use *Foursquare* APIs to discover as many venues as possible in this region and we collect the check-ins at these venues. We have removed the venues with only one check-in. Such single check-ins are not useful for topic categorization and may introduce noise in the LDA model that we use. Our data set is summarized in Table 1.

Foursquare defines a hierarchical list of categories applied to venues. There are 9 top categories in the hierarchical structure and they are: *Arts & Entertainment*, *College & University*, *Food*, *Professional & Other Places*, *Nightlife Spot*, *Great Outdoors*, *Shop & Service*, *Travel & Transport* and *Residence*. In our data set, there are 45,125 check-ins at the venues that do not belong to any category. Figure 1 shows the venue distribution of the other 768,096 check-ins in the top 9 categories.

From Figure 1, we can see that the check-ins in *Food* and *Shop & Service* categories are always the most for weekdays, weekends as well as overall. However, the check-ins in *College & University* and *Professional & Others* categories on weekdays are far more than those at weekends. This is not difficult to understand as users usually do not work at

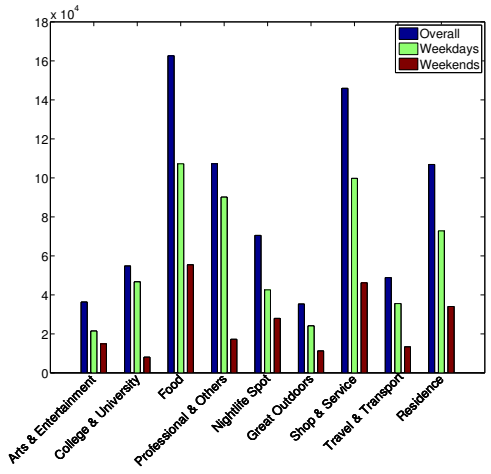

Figure 1. Check-ins distribution in top 9 categories

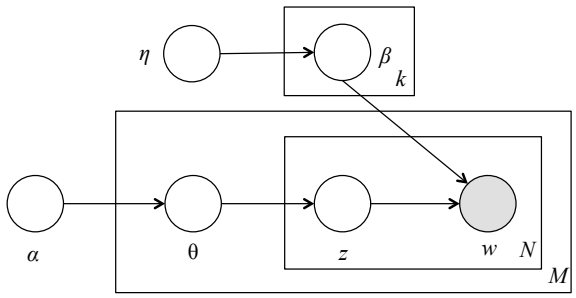

Figure 2. Graphic model representation of LDA [6]. The boxes are "plates" representing replicates. The plate M represents documents, while the plate N represents the repeated choice of topics k and words within a document.

weekends. These categories can help to understand the local geographic topics as described in the experiment section.

GEOGRAPHIC TOPIC MODELING

In this section, we first introduce the LDA model and then show how to use it to discover the geographic topics.

LDA Model

Blei *et al.* present Latent Dirichlet Allocation (LDA), which is a probabilistic model in [6]. LDA is usually used to cluster documents based on the topics contained in a corpus of documents. Ferrari and Mamei present two reasons to choose the LDA in analyzing users' mobility patterns and routine behaviors in [13]. One is that there is no need to define topics a priori and the other is that the topic results represent meaningful probabilistic distributions over words and documents [13].

The graphic model of LDA is shown in Figure 2. α and η are parameters of the Dirichlet prior on the per-document topic distributions and per-topic word distributions, respectively. θ_i is the topic distribution for document i. β_k is the word distribution for topic k. w_{ij} is the jth word in document i and z_{ij} is the topic for w_{ij}.

Trajectory-driven Geographic Topic Modeling

We adopt the LDA to identify the geographic topics in our data set. The basic unit is word in the LDA and, in this paper, the venue in a single check-in record represents a word. A user's trajectory consisting of all the venues of his check-ins represents a document, which is a set of words. An advantage of using LDA is that we do not need to predefine the topics and only need to set the number of the topics.

We focus on the local geographic topics in this paper; so M users' trajectories within a specific city make up the corpus in the model. Every user's trajectory can be described as a mixture of the geographic topics that are essentially distributions over the geo-locations. Therefore, the LDA can be applied to the mobility data in this way.

EXPERIMENT

In this section, we first introduce our data set and describe how we process the data set to make it suitable for the LDA model. After that, we present several experiments to evaluate the proposed approach. In particular, we first investigate the overall local geographic topics in our data set. Since the users' check-ins patterns differ for weekdays and weekends [8, 12], we also explore the geographic topics on weekdays and at weekends. We use the MATLAB Topic Modeling Toolbox to run our experiments[16].

Data Preparation

In each check-in record, we have the user who created the check-in and the venue where the check-in was created. Moreover, we can get the creation timestamp of the check-in. We conduct three experiments. In the first experiment, we do not consider the creation timestamps of the check-ins. In the second and third experiments, we divide our data set into two subsets according to the creation timestamps of the check-ins, i.e. the weekday subset consists of the check-ins created on weekdays and the weekend subset consists of the check-ins created at weekends.

In these three experiments, the basic units in the trajectory-driven geographic topic model are the venues of the check-ins. A document consists of the venues from a user's check-ins, e.g. ($venue_{check-in1}$,...,$venue_{check-inN}$). We set the number of topics $T = 30$, $\alpha = 50/T$ and $\eta = 0.1$ in all these experiments.

Local Geographic Topics

In this subsection, we present the overall local topics discovered by the first experiment where the timestamps of the check-ins are not counted. We get a total of 30 local geographic topics and use OTopics to denote the topics in the first experiment. Table 2 lists the 8 OTopics derived from the first experiment. In each OTopic, we also list top 10 venues. In addition, we discuss the spatial features in these topics.

Overall Local Geographic Topics

In the section where we introduce our data set, we plot the distribution of check-ins in top 9 categories. From Figure 1 we can see that the check-ins in *Food* and *Shop & Service* are the largest in number. In the generated OTopics, we find that many of them are related to these two categories. For instance, OTopic 7, 9, 19 and 28 are topics related to food and shopping and most of the top 10 venues in these four topics are food and shopping venues. It indicates that users would go to restaurants after shopping. The different topics give an overview of clusters of the shops and restaurants people usually check in at together.

We also have OTopics that are related to education. OTopic 3 is associated with the University of Pittsburgh and OTopic 5 is associated with the Carnegie Mellon University. Both the universities are located at Oakland neighborhood in Pittsburgh. The food venues in these two topics indicate places that students and faculty members most likely go frequently.

Some OTopics are related to sports. OTopic 4 is a case in point. CONSOL Energy Center is the home stadium of the hockey team–the Pittsburgh Penguins and PNC Park is the home stadium of the baseball team–the Pittsburgh Pirates. Moreover, the probabilities of these two venues assigned to this topic are ten times higher than the other venues in this topic. There is also a very famous football team in Pittsburgh–the Pittsburgh Steelers, whose home stadium is Heinz Field. The reason that Heinz Field is not in this topic is because our data collection period does not overlap with the football season.

There are also OTopics that are related to businesses such as OTopic 11, since there are several professional buildings in this topic. An interesting observation is that there are two tunnels in this topic, which are located in the major route of Pittsburgh (I–376). The existence of these two venues indicates that many users commute between work and home through these tunnels.

In summary, the advantages of the user-driven local geographic topics include the following: (1) the topics generated depend only on the users' trajectories but not on the physical locations or the pre-defined categories; and (2) the topics provide a novel view of the location classifications based on the human mobility. That is, the venues in a geographic topic imply that people usually go there together with high probability.

The Spatial Features of the Topics

We are also interested in the spatial features of the generated local geographic topics, i.e. whether the venues in each topic are close to each other or not? Whether there are any spatial relations among the venues in the same topic?

Figure 3 shows the spatial distribution of the top 10 venues in the eight OTopics illustrated in Table 2. We can see that the venues in some OTopics are very close, e.g. the top 10 venues in OTopic 5. The venues in some other OTopics are very sparse, e.g. the top 10 venues in OTopic 4. Considering the categories of the topics, venues in the topics related to education (OTopic 3 and OTopic 5) are usually geographically closer. Venues in topics related to businesses or entertainment may not be close to each other. It is not difficult to understand because people usually commute between home

Topic 7	Topic 9	Topic 19	Topic 28
South Hills Village Mall	AMC Loews Waterfront 22	Galleria at Pittsburgh Mills	Walmart Supercenter
Giant Eagle Market District	P.F. Chang's	Cinemark IMAX Theater	Quaker Steak & Lube
Starbucks	Target	Walmart Supercenter	Buffalo Wild Wings
Red Robin Gourmet Burgers	Giant Eagle	Do Drop Inn	Primanti Brothers
Houlihan's Mt. Lebanon	Planet Fitness	Giant Eagle	Costco
T.G.I Friday's	Red Robin Gourmet Burgers	Walmart Supercenter	Giant Eagle Market District
Trader Joe's	Eat'n Park	Target	Giant Eagle
Giant Eagle	T.G.I. Friday's	Applebee's	Starbucks
Walmart	Costco	Giant Eagle	North Park Lounge in HD
GetGo	The Waterfront	UPMC St. Margaret Hospital	Target
Topic 3	**Topic 5**	**Topic 4**	**Topic 11**
Cathedral of Learning	University Center	CONSOL Energy Center	BNY Mellon Center
Hillman Library	Gates-Hillman Complex	PNC Park	Comcast
Hemingway's Cafe	USX Tower	PNC Park	American Eagle HQ
Benedum Hall	Wean Hall	Olive Garden	Fort Pitt Tunnel
Posvar Hall	Hunt Library	Starbucks	Fitness 247
Petersen Events Center	Morewood Gardens	Joe's Crab Shack	PNC Firstside Center
William Pitt Union	Starbucks	Verizon Wireless	"Bellevue, PA"
Peter's Pub	Panther Hollow Inn	St. Clair Hospital	Squirrel Hill Tunnel
Schenley Plaza	Doherty Hall	CCAC Milton Hall	Crawford Square Apartments & Townhomes
Chipotle Mexican Grill	Hamburg Hall	U.S. Steel Clairton Works	Element Church 205 North

Table 2. Examples of Trajectory-driven Overall Local Geographic Topics

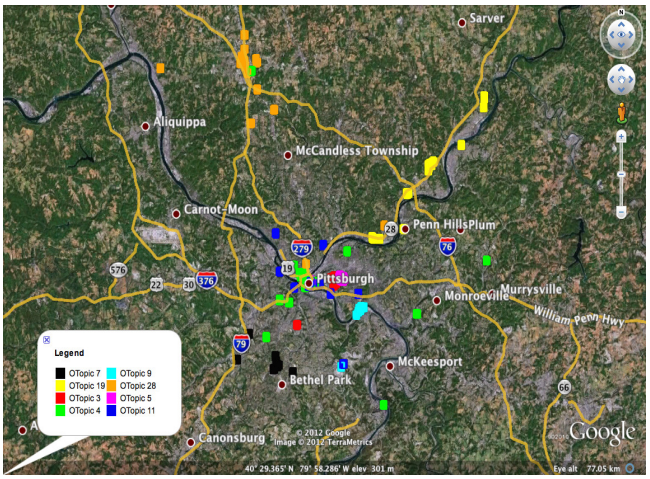

Figure 3. The spatial features of the topics

and work (e.g. OTopic 11) and the stadiums are usually not close to the fans' home (e.g. OTopic 4). However, the case is different for the OTopics related to shopping. Some people would like to go shopping and dining at venues that are not far away. For example, OTopic 9 and OTopic 7; the top 10 venues in these two topics are very close. Some people would like to go shopping and dining at venues that are along the freeway, e.g. OTopic 19; the top 10 venues in this topic are along the freeway.

Figure 3 essentially indicates that venues in the user trajectory-driven local geographic topics are not always geographically close to each other. Our work is a strong complement to the recent work based on the topic model or clustering of the venues in LBSNSs presented in [8, 12]. Moreover, our work can be used for business planning. For example, if there is a store or restaurant that is not close to the majority of top venues in a topic, the owner of the store or restaurant may need to consider if a branch office should be opened in the area with the majority of venues in the topic.

Local Geographic Topics on Weekdays vs. Those at Weekends

We analyze the geographic topics on weekdays and at weekends in the second and third experiments in this subsection as a complementary for Figure 1 that show the differences between the weekday check-ins and the weekend check-ins.

Local Geographic Topics on Weekdays

We expect that there would be local geographic topics related to university, professional work and business on weekdays and we use WDTopics to denote the topics on weekdays. The results confirm our expectation. We list some topics in Table 3. WDTopic 25 relates to the University of Pittsburgh and WDTopic 26 relates to the Carnegie Mellon University. The top 10 venues in these two topics are almost the same with those in OTopic 3 and OTopic 5.

We have several WDTopics related to professional work and business. For instance, WDTopic 5 is related to medical work, as 8 out of 10 top venues in this topic are hospitals. One possible reason could be that patients usually go to

WDTopic 25	WDTopic 26	WDTopic 5	WDTopic 17
Cathedral of Learning	University Center	UPMC Presbyterian Hospital	Rivers Casino
Hillman Library	Gates-Hillman Complex	UPMC Shadyside	IKEA
Benedum Hall	Wean Hall	Allegheny General Hospital	American Eagle HQ
Hemingway's Cafe	Hunt Library	UPMC Mercy	Comcast
Posvar Hall	BNY Mellon Client Service Center	UPMC Montefiore	Emerald Gardens Apartments
William Pitt Union	Morewood Gardens	UPMC St. Margaret Hospital	Oakmont Tavern
David Lawrence Hall	Porter Hall	Verizon Wireless	Fort Pitt Tunnel
Schenley Plaza	Doherty Hall	Western Pennsylvania Hospital	Rivertowne North Shore
Petersen Events Center	Starbucks	Forbes Regional Hospital	HSLS Falk Library: 200 Scaife Hall
Mad Mex	Tepper School of Business	"Plum, PA"	Squirrel Hill Tunnel
WDTopic 1	**WDTopic 18**	**WDTopic 7**	**WDTopic 9**
CONSOL Energy Center	PNC Park	Pittsburgh International Airport (PIT)	AMC Loews Waterfront 22
Urban Active Fitness	Stage AE	Pittsburgh International Arprt	P.F. Chang's
PNC YMCA	Heinz Hall	The Westin Convention Center	Walmart Supercenter
PITT School Of Information Sciences	Forbes Tower	David L. Lawrence Convention Center	Target
Buffalo Wild Wings	Hamburg Hall	Wyndham Grand Pittsburgh Downtown	Century III Mall
DoubleTree - Green Tree	Heinz Field	T.G.I. Friday's	Giant Eagle
Oakland	PNC Park	Freedom High School	Red Robin Gourmet Burgers
Bear Run Village	Olive Garden	Heritage Hills Townhomes Apartments	Eat'n Park
Renaissance Pittsburgh Hotel	Starbucks	EFI, Inc. (Pittsburgh Office)	Amberson Towers
Ariba Inc.	Pittsburgh International Airport (PIT)	Tonic Bar And Grill	Jefferson Regional Medical Center

Table 3. Examples of Trajectory-driven Geographic Topics on Weekdays

these hospitals for different purposes. Since the University of Pittsburgh Medical Center (UPMC) is the largest medical center in western Pennsylvania area, patients may have been referred to different specialists who work in different hospitals belonging to the UPMC. Another reason could be that the doctors, nurses and other medical staff work at different locations at different times. We do not identify a topic related to the medical work in the overall topics.

We also have a topic (WDTopic 17) with two tunnels–Fort Pitt Tunnel and Squirrel Hill Tunnel that are also included in OTopic 11. However, the top 10 venues in WDTopic 17 are a little different from those in OTopic 11. Rivers Casino and IKEA are in WDTopic 17 but not in OTopic 11. The reason may be that there are many people who work at Rivers Casino and IKEA and usually commute on the main route I–376 on weekdays.

The WDTopics related to sports are a little different with the OTopic related to sports. We can see that CONSOL Energy Center is in WDTopic 1 and PNC Park is in WDTopic 18. In both these topics there are hotels in the top 10 venues that may imply that hockey fans or baseball fans fly to Pittsburgh for a game on weekdays. We also see University of Pittsburgh School of Information Sciences and Oakland in WDTopic 1. It may be because university students are more likely to watch the early hokey game on weekdays at a very cheap price [4].

WDTopic 7 shows that airport on weekdays are related to businesses, as there are many conferences or meetings at The Westin Convention Center, David L. Lawrence Convention Center and Wyndham Grand Pittsburgh Downtown.

WDTopic 9 gives an example of the topics in weekdays related to shopping and food. The venues in WDTopic 9 are almost the same as the venues in OTopic 9.

Local Geographic Topics at Weekends

At weekends, we can expect more topics related to entertainment and shopping. Besides, there would be very few topics related to education, business and professional work. Our results indeed confirm these. In the 30 topics, there is no topic related to any university, which indicates that there are not as

WETopic 4	WETopic 16	WETopic 10	WETopic 27
Pittsburgh International Airport (PIT)	Heinz Hall	Target	The Cheesecake Factory
Pittsburgh International Arprt	Carnegie Science Center	Eat'n Park	Bar Louie
Hard Rock Cafe Pittsburgh	Pittsburgh Zoo & PPG Aquarium	Giant Eagle	Claddagh Irish Pub
Robert Morris University Island Sports Center	Petersen Events Center	The Waterfront	Pittsburgh Marriott City Center
T.G.I. Friday's	Red Robin Gourmet Burgers	Dave & Buster's	Emerald Gardens Apartments
3:36	Carnegie Museum of Natural History	T.G.I. Friday's	The Altar Bar
Baggage Claim	Joe's Crab Shack	AMC Loews Waterfront 22	Grand Concourse
Hard Rock Cafe	Station Square	Barnes & Noble	Megabus Pittsburgh
"Freedom, Pa"	Bar Room	Costco	Starbucks
House	Eat'n Park	Giant Eagle	Joe's Crab Shack

Table 4. Examples of Trajectory-driven Geographic Topics at Weekends

many users at weekends checking in at universities as those on weekdays in *Foursquare*. There is no topic about professional work or business, either. All the topics are related to food, shopping and entertainment. We use WETopics to denote the local geographic topics at weekends and Table 4 lists some examples.

WETopic 4 shows the geographic topic related to airport. We do not see any convention centers in the top 10 venues in this topic. It may imply that users usually do not take flight for business reasons at weekends. WETopic 16 is mainly related to arts and entertainment. There are typically many concerts, operas and stage shows in Heinz Hall. Carnegie Science Center and Carnegie Museum of Natural History are very famous museums in Pittsburgh. Petersen Center usually host university sport events. WETopic 10 is related to shopping and food. The top 10 venues in this topic are very similar with those in OTopic 9. WETopic 25 is mainly related to food, which is different from the OTopics as the food venues usually appear together with the venues in shopping or the venues in entertainment.

Comparisons Between WDTopics and WETopics

Comparing all the local geographic topics on weekdays and those at weekends, we observe the following:

- There is no topic related to education, business or professional work out of the 30 WETopics. This is not difficult to understand as people usually do not go to school or work at weekends; rather they most likely engage in social, family events and recreations at weekends. Thus, the topics at weekends are mostly about entertainment and recreations. We also have topics about shopping and various entertainment venues, which are more than those on weekdays.
- There are topics related to food at weekends. However, there is no such a topic on weekdays. Food venues usually appear together with shopping, entertainment or business venues in the local geographic topics on weekdays.
- A few art and science centers appear in the topics on weekdays but they show up as a majority in several topics at weekends. One possible reason could be that these art and science centers are usually closed after 5pm on weekdays, thus people may have no time to visit such centers after the work. Another possible reason could be that it usually takes a long time to visit such museums and art centers so people would like to visit them on weekends.

In summary, the differences between the local geographic topics on weekdays and those at weekends correspond with the differences of the human mobility on weekdays and at weekends. Therefore, our proposed approach can characterize the human mobility patterns.

DISCUSSION

Interesting Observations

From the three experiments in the last section, we observe several interesting findings.

Hospitals & Fitness Centers In our OTopics, we find that hospitals and fitness centers appear in the same topics frequently. There are 9 hospitals and 11 fitness centers in the top 10 venues of the 30 OTopics. 5 hospitals and 6 fitness centers appear together in 5 different topics. One reason of the co-appearances of the hospitals and fitness centers may be that medical workers do exercises frequently. It also may be that patients often go to fitness centers to improve health.

Topics involving Airports We have one topic related to airport on weekdays (WDTopic 7) and one topic related to airport at weekends (WETopic 4). We find significant differences between the other venues of the top 10 venues in these two topics. On weekdays, venues are mainly related to business. For example, The Westin Convention Center and David L. Lawrence Convention Center in WDTopic 7 are big convention centers for conferences. We do not see such venues in WETopic 4. This difference also complies with the human mobility pattern as there are many people taking flights for business purposes on weekdays.

Local Supermarkets vs. Global Supermarkets Giant Eagle

is a supermarket chain in Pittsburgh and Walmart is a global supermarket chain. Both of these supermarkets provide very similar groceries and services. However, they are often in the same topics indicating that perhaps many people go to both of them.

Local Coffee Shop Chain vs. Global Coffee Shop Chain Crazy Mocha is a famous local coffee shop chain in Pittsburgh and Starbucks is a very famous global coffee shop chain. In our local geographic topics, Starbucks appears to be one of the top 10 venues in many topics, while Crazy Mocha is not included by one of them. In our data set, the number of Starbucks is almost 4 times the number of Crazy Mocha; however, the number of check-ins at Starbucks is ten times more than that at Crazy Mocha. Thus, it seems that the customers do not check in at Crazy Mocha as often as they do at Starbucks.

Applications

Next, we also discuss how our proposed approach and results can be used in different applications.

Friend Recommendation Our proposed topic model based approach can be used in friend recommendation. Since each user's trajectory can be described by the topics, the similarity of the topics could be helpful in recommending friends to users. For example, for two PITT students who check in at Pitt's Hillman Library, School of Information Sciences and William Pitt Union frequently, although they may not know each other right now, they may register in the same course, or join the same student organization or interest group in the future, and the probability of being friends will be very high.

Location Recommendation & Prediction Our proposed topic model based approach can also be used in location recommendation. Our proposed topic model is based on the users' trajectories, thus the venues in the same topic are the ones that many people usually go to. Personalized location recommendation should also consider such venues as candidates for recommendations because they are more or less "hot spots" for a group of people. For example, some parents would like to take their children to visit the museums or science centers at weekends. Thus, the topics about entertainment could provide very valuable references since many people have been there. The topic model also can be used in location prediction. If some similar users or friends of a user have been to a very good venue, the user may also go to the venue with high probability. In the case of the entertainment example above, the venues in the entertainment topic may be the next check-in venues for the parents and their children. Even without the topic model based recommendation, the parents still will go to the museums or science centers they have never been to before, as they would likely bring their kids there. Our proposed topic model based approach can still help to predict locations as it can characterize human mobility pattern.

Business Strategy Design Users' trajectory-driven local geographic topics can be helpful in designing business strategies. The topics can help a business owner to find out whether there is any complementary relationship between his venues and other venues and then help him explore the potential locations for a new chain store. It is also helpful for the business owners to identify his competitors' advantages to improve his own business. For example, it could be interesting for the Crazy Mocha to figure out the reasons why it has far fewer check-ins compared to that of Starbucks.

RELATED WORK

Li and Chen present their work of large-scale quantitative analysis of LBSNS in [15]. Their work is very general in analyzing the user profiles, update activities, mobility characteristics, social graphs and attribute correlations.

Cheng *et al*. explore the check-ins to analyze human mobility patterns in the spatial, temporal, social, and textual aspects [17]. Their work uses the global data collected from *Twitter*, which is different from ours, as our data set are local data set from *Foursquare*. Moreover, they do not use topic models to analyze the check-ins.

Noulas *et al*. study the user behavior in *Foursquare* in [5]. They investigate the check-in dynamics in spatio-temporal aspects. However, they do not analyze the relations among the check-ins.

Ferrari *et al*. employ LDA to extract the urban patterns from location-based social networks in [12]. Our work is different from theirs mainly in two ways. First, our topics are based on the users' trajectories, i.e., we use a user's trajectory as a document and a venue in a check-in as a word in the LDA model. However, their work uses a venue and a time slot as a word and a day of the city is a document in their LDA model. Thus, their work focuses on the human mobility pattern during different times within a city and our work is more human centric than theirs as we investigate the topics based on trajectories of large groups of users. The data set used in their work is crawled from *Twitter* but not from *Foursquare* directly. In *Foursquare*, a user can use his *Twitter* account to login and post his check-ins on *Twitter*, but not every user has a *Twitter* account and neither is every user likely to post his check-ins on *Twitter*. Thus, our data is more comprehensive and complete than the *Twitter* data set.

Ferrari and Mamei also investigate the topics based on the user's trajectory in [13]. The data set in [13] is the daily whereabouts of two persons over the period for almost one year. They divide a day into 48 time slots and each time slot lasts for 30 minutes, so the 48 places each day form a document. Thus, their work in [13] is still time based topics, which focuses on a single user's mobility pattern at different times and is thus very different than our approach.

Farrahi and Gatica-Perez's work in [11] also use LDA to discover the routine behaviors. They use the Reality Mining data set [14] from MIT that contains a one-year mobile phone sensor data recording 97 subjects from 2004 to 2005. The routine behaviors in their work are still temporal based topics, which are different from ours, as we do not consider the temporal factors in our model. Besides, the locations in

their work are simply labeled by "Home", "Work", "Other" and "No Reception". Thus, the rich venue information is lost in their work.

Yuan *et al*. also propose a framework to discover regions of different functions in a city by using human mobility among both regions and points of interests in the region in [10]. Their topic model is based on LDA and Dirichlet Multinomial Regression. In their work, they use the GPS trajectory datasets. Besides, their work aims to discover the region topics, which is different from ours.

Cranshaw and Yano employ LDA to distill the proto-neighborhoods from *Foursquare* data set in [9]. In their work, the word is the category of the venue and the document is the check-ins in a region. Regions are small grids that divide space according to the latitude and longitude space. Thus, each region can be described by the topics, which can help to understand the neighborhood.

Chang and Sun analyze users' check-ins in general on *Facebook Place* in [7] and LDA is also used in their work to investigate the user membership in a low-dimensional representation of the place space. Since their work is in general about the check-in analysis so the topic model is only a small part and they only give three topics without analyzing the topics in details as we do. Besides, they do not consider the differences in mobility patterns of users in the weekdays and weekends.

CONCLUSION AND FUTURE WORK

In this paper, we have employed the LDA model to investigate the local geographic topics based on the users' check-ins in *Foursquare*. Since our topics are derived from the trajectories of 32,113 different users, the local geographic topics in this paper indicate the co-check-in among such users. That is, the venues in the same topics are usually co-occurrences in many users' trajectories. The analysis of the local geographic topics also verifies the effectiveness of the model. For example, we can see there are topics describing universities, entertainment venues, etc. Besides, we also explore the spatial features of the top 10 venues in some topics and we find that the venues in the same topic can be either close or far away from each other. Thus, the topics are not limited by the spatial information. Moreover, we study the local geographic topics on weekdays and at weekends, respectively. The differences between them comply with the human routine behaviors; the results also demonstrate the effectiveness of our approach. Furthermore, we discuss several interesting findings of our topics and the applications of the proposed topic model based approach.

One future research direction is to use the topic model in location and friend recommendations. We plan to also investigate approaches other than LDA in discovering the local geographic topics from the large-scale LBSNS data set.

ACKNOWLEDGMENTS

This research has been supported by the US National Science Foundation award IIS-0545912. We would like to thank Mark Steyvers and Tom Griths for providing their MATLAB Topic Modeling Toolbox which is used in this paper.

REFERENCES

1. Zheng, Y. Location-based social networks: Users. *Computing with Spatial Trajectories*, Zheng, Y and Zhou, X., Eds. Springer, 2011.
2. Foursquare Nears 20 Million Users And Crowley Talks About His Co-founder's Recent Departure. `http://articles.businessinsider.com/2012-03-10/tech/31142426_1_foursquare-sxsw-dennis-crowley`.
3. 200M users include location in Facebook posts; company looks to expand location APIs. `http://www.insidefacebook.com/2012/04/05/200m-users-include-location-in-facebook-posts-company-looks-to-expand-location-apis/`.
4. Get tickets for select penguins home games with american eagle student rush. `http://www.ticketmaster.com/promo/u3c7b3?brand=penguins`.
5. Noulas, A., Scellato, S., Mascolo, C. and Pontil, M. Empirical study of geographic user activity patterns in foursquare. In *ICWSM'11*, 2011.
6. Blei, D.M., Ng, A.Y., and Jordan, M.I. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
7. Chang, J. and Sun, E. Location3: How users share and respond to location-based data on social. In *ICWSM'11*, 2011.
8. Cranshaw, J., Schwartz, R., Hong, J.I. and Sadeh, N. The livehoods project: Utilizing social media to understand the dynamics of a city. In *ICWSM'12*, 2012.
9. Cranshaw, J. and Yano, T. Seeing a home away from the home: Distilling proto-neighborhoods from incidental data with latent topic modeling. In *CSSWC Workshop at NIPS'10*, 2010.
10. Yuan, J., Zheng, Y. and Xie, X. Discovering regions of different functions in a city using human mobility and pois. In *KDD'12*, 2012.
11. Farrahi, K. and Gatica-Perez, D. Discovering routines from large-scale human locations using probabilistic topic models. *ACM Transactions on Intelligent Systems and Technology*, 2(1), 2011.
12. Ferrari, L., Rosi, A., Mamei, M. and Zambonelli, F. Extracting urban patterns from location-based social networks. In *LBSN'11*, 2011.
13. Ferrari, L. and Mamei, M. Discovering daily routines from google latitude with topic models. In *CoMoRea'11*, 2011.
14. Eagle, N., Pentland, A. and Lazer, D. Inferring social network structure using mobile phone data. In *Proceedings of the National Academy of Sciences (PNAS)*, volume 106, pages 15274–15278, 2009.
15. Li, N. and Chen, G. Analysis of a location-based social network. In *CSE'09*, 2009.
16. Griffiths, T. and Steyvers, M. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(supple.1):5228–5235, 2004.
17. Cheng, Z., Caverlee, J., Lee, K. and Sui, D.Z. Exploring millions of footprints in location sharing services. In *ICWSM'11*, 2011.