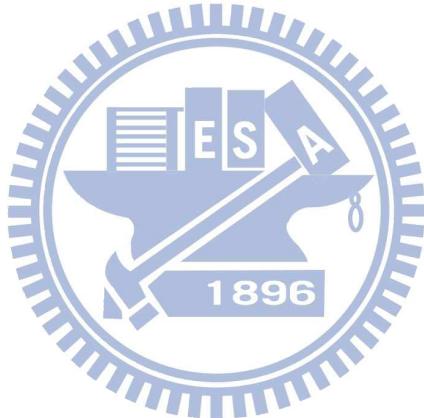


Contents

1	Introduction	1
2	Preliminaries	5
2.1	Notations and Problem Definitions	5
2.2	Framework	6
3	Activity and Location Inference in LBSNs	8
3.1	Relations among Time, Location and Activity	8
3.2	Activity Inference, Deriving $P(a u, \ell, t)$	9
3.3	Mobility Inference, Deriving $P(\ell a, t, u)$	10
3.4	Time-Activity Model, $P(a t, u)$	10
3.4.1	Activity Transition Model	12
3.5	Location-Activity Model, $P(\ell a, u)$	13
3.6	Performing Activity and Mobility Inference	14
3.6.1	Constructing Time-Activity Model	14
3.6.2	Constructing Location-Activity Model	15
3.6.3	Performing Activity Inference	16
3.6.4	Performing Mobility Inference	16

4	Performance Evaluation	17
4.1	Datasets Description	17
4.2	Impact of Time Slot Length	18
4.3	Comparison of Activity Inference Methods	19
4.4	Comparison of Location Inference Methods	20
5	Related Works	22
6	Conclusion	24

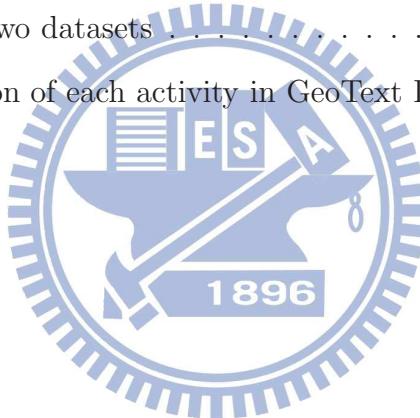


List of Figures

1.1	User's check-in records and activities in an LBSN	3
2.1	The framework for activity and mobility influence	7
3.1	The Bayesian network structure describes the relation among time, activity and location factors	9
3.2	Temporal distribution over time, where the gray bar shows the frequency of each activity	11
4.1	Impact of time slot length and the training size	19
4.2	Comparison of activity inference methods	20
4.3	Comparison of location inference methods	21

List of Tables

1	Statistics of two datasets	18
2	The proportion of each activity in GeoText Dataset	18



Chapter 1

Introduction



With the increasing of location-aware techniques equipped mobile devices, many location-based social networks (LBSNs) provide users to check-in at a location to share to their friends for more social interactions, such as Facebook¹, Twitter² and Foursquare³. When users check-in at a location, they can leave some text information in their check-ins, such as tags and text messages. The text information sometimes describe what is the user doing when he/she checks-in at a location. Thus, the check-in message contains not only the spatial-temporal information but also the text information.

To provide personalized location-based services in LBSNs, such as coupons and ads, two problems have to be solved. One is to know whether a user performs the given activity at a specific time and location. For example, a restaurant would like to find who will have lunch or dinner near by the restaurant, and it has higher probability to promote the meal coupon to him

¹<https://www.facebook.com/>

²<https://twitter.com/>

³<https://foursquare.com/>

or her successfully. The other is to know where a user may perform the given activity at a specific time. For example, a restaurant chain would like to know where a user will have lunch or dinner, and promotes the near restaurants to the user. Figure 1.1 shows an example. The user has two check-in records at ℓ_1 and ℓ_2 at t_1 and t_2 , respectively. If a coupon is available at t_3 at ℓ_3 , the shop will decide whether promote the user if the user activity matches the type of shop, such as dining and restaurant. Furthermore, a shopping center chain want to know where is the user at t_4 , and then it can promote near shop to this user for shopping.

Prior works [19] utilize Bayesian network to model the relations among user, location, time and activity in Twitter. In [19], the network structure consists of user, location, region, time, weekend/weekday, words and topics. However, [19] only counts and extracts the representative words. It is hard to extract the same activities in different languages with different characters. To more precisely capture user activities, [15] clusters user labels to activity labels. Some natural language processing works focus on extracting topics or activities from text, such as Latent Dirichlet allocation (LDA) topic model [16]. Moreover, crowdsourcing is another way to extract activities from text precisely [22]. Thus, we assume the activities are already extracted from other precise methods to against different language text in LBSNs. To the best of our knowledge, in LBSNs, prior works focus on discovering spatial-temporal relations [13][2][18] or geographical topics [7][20]. Furthermore, in [17], the authors try to discover tags of locations in an LBSN. However, the tags of locations are for global users. It can not reflect individual location preference. For example, there are sport field, food court and office in a college. Thus, we

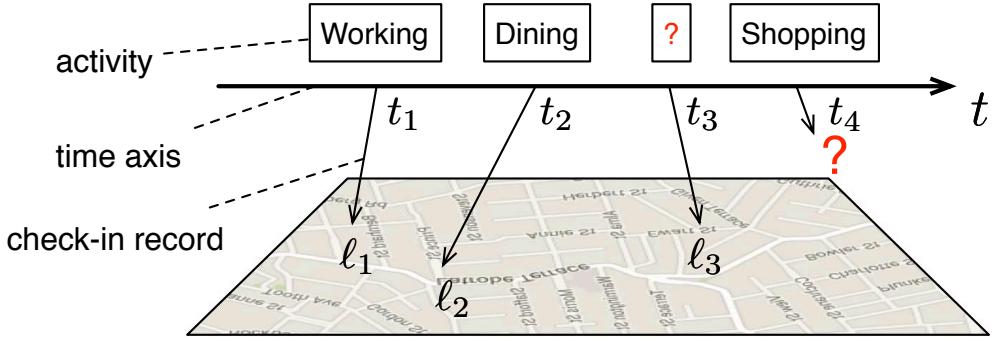


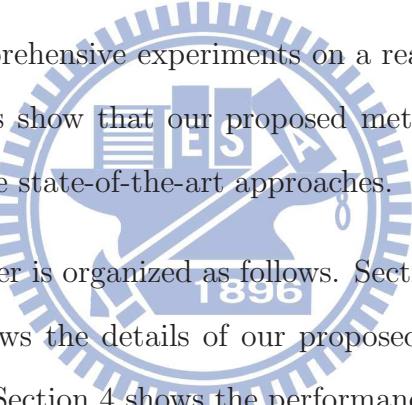
Figure 1.1: User’s check-in records and activities in an LBSN

will take the spatial, temporal and activity factor into account for individual activity and mobility influence problem.

In this paper, we aim to infer individual activity and mobility based on their check-in records in LBSNs. Our idea is to utilize Bayesian theorem [6] to evaluate the probability of user u performing an activity a at location ℓ at time t and the location distribution of user u performing an activity a at time t . To evaluate the distributions, we propose a Bayesian-based approach to simplify the probability equation and get the approximation. This approach exploits Bayesian network structure which consists of the relations among time, location and activity factors. After the simplified equation, there are two major components we have to deal with. One is the relation between location and activity. The other is the relation between activity and time. First, to infer the mobility based on activity, we utilize Gaussian mixture model (GMM) to capture spatial preference of user with different activities. Second, to infer the activity at a specific time, we propose the Activity Transition Model to infer user activity at a specific time.

In summary, our major contributions are outlined as follows:

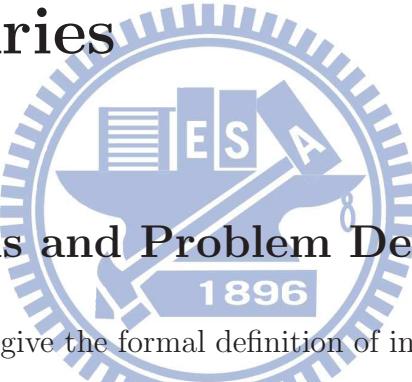
- We formulate the individual activity and mobility inference problem for personalized location-based services in LBSNs.
- We utilize Bayesian network to describe the relations among spatial, temporal and activity factors.
- We proposed a Bayesian-based approach for activity and mobility influence in LBSNs based on our proposed Bayesian network structure.
- We conduct comprehensive experiments on a real dataset, and the experimental results show that our proposed methods have higher performance than the state-of-the-art approaches.



The rest of this paper is organized as follows. Section 2 defines the problem, and Section 3 shows the details of our proposed methods for activity inference in an LBSN. Section 4 shows the performance evaluation on a real dataset. Section 5 gives an overview of related works. Finally, Section 6 concludes this paper.

Chapter 2

Preliminaries



2.1 Notations and Problem Definitions

In this section, we will give the formal definition of individual activity inference problem in LBSNs. First, we define the activities we concerned in this paper. The details are as follows.

Definition 1 Categories/Activities

There are ten types of categories from Foursquare we concerned in this work, which are Arts & Entertainment (AE), Colleges & Universities (CU), Events (E), Food (F), Nightlife Spots (NS), Outdoors & Recreation (OR), Professional & Other Places (PO), Residences (R), Shops & Services (SS), and Travel & Transport (TT). We regard the ten categories as user activities in this paper. Hence, the activity set $A = \{AE, CU, E, F, NS, OR, PO, R, SS, TT\}$.

Then, the definition of check-in records in an LBSN is as follows.

Definition 2 Check-in Records in an LBSN

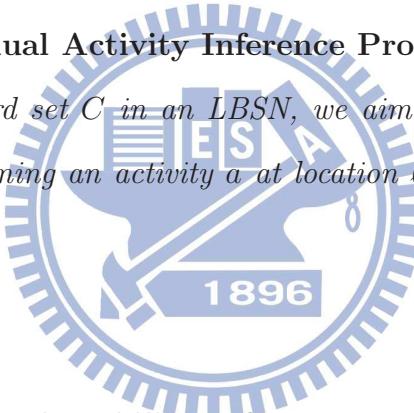
The set of check-in records in an LBSN $C = \{(u, a, \ell, t)\}$, where (u, a, ℓ, t) denotes a check-in record that a user u performs activity a at location ℓ at time t , and $\ell \in L$. A location ℓ is a coordinate which consists of latitude and longitude.

Finally, the definition of our target problems in this paper are as follows:

Definition 3 Individual Activity Inference Problem

Given a check-in record set C in an LBSN, we aim to evaluate the probability of user u performing an activity a at location ℓ at time t , denoted by $P(a|u, \ell, t)$.

and



Definition 4 Individual Mobility Inference Problem

Given a check-in record set C in an LBSN, we aim to find the location distribution of user u performing an activity a at time t , denoted by $P(\ell|u, a, t)$.

2.2 Framework

Figure 2.1 shows the framework for individual activity and mobility inference in LBSNs. The input check-in records contain the information of user, time, location and activity. In the offline stage, the two major component models, time-activity and location-activity model, capture the behavior among location, time and activity factors. In the online stage, our framework accepts two types of queries, individual activity and mobility inference. The two

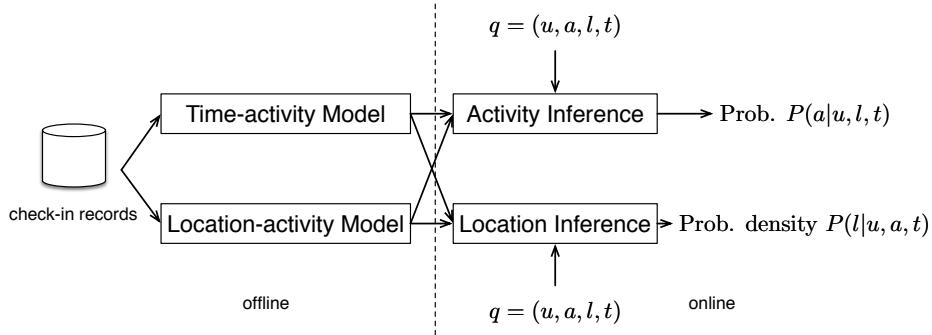
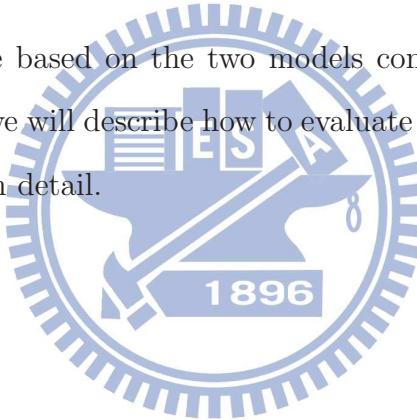


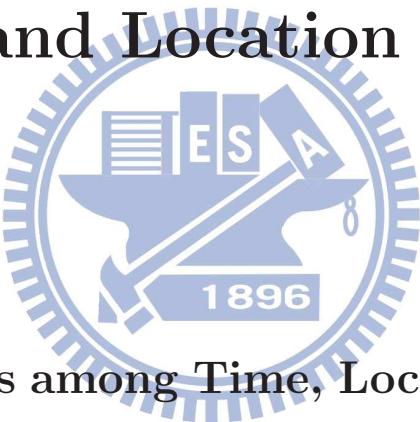
Figure 2.1: The framework for activity and mobility influence

inference functions are based on the two models constructed in offline. In the following section, we will describe how to evaluate $P(a|u, \ell, t)$, $P(\ell|u, a, t)$ and each component in detail.



Chapter 3

Activity and Location Inference in LBSNs



3.1 Relations among Time, Location and Activity

As we know, the mobility of users in LBSNs has regularity property [2], i.e. users may visit a location at a specific time. Users visit locations due to not only the time but also the activity. For example, a user goes to his office every morning, visits a cafeteria at noon, and then drives home in the evening. It seems like that user locations depend on the temporal feature. But if we investigate further, the user goes to his office and a cafeteria regardless of the time, but the activities, which are working and dining, respectively. In other words, the activity effect on location is much stronger than the temporal effect does. In consequence, we assume the time factor as the factor that

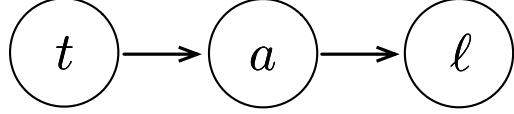


Figure 3.1: The Bayesian network structure describes the relation among time, activity and location factors

influences user activities, while the user activities effects the geographical feature. Figure 3.1 shows the Bayesian network structure among these three factors based on above idea. The activity factor is related to the time factor, and the location factor is related to the activity factor.

3.2 Activity Inference, Deriving $P(a|u, \ell, t)$

Based on the Bayesian network structure mentioned in Section 3.1, given user u , location ℓ and time t , the activity inference, $P(a|u, \ell, t)$, can be simply as following steps:

$$\begin{aligned}
P(a|u, \ell, t) &= \frac{P(a, \ell, t|u)}{P(\ell, t|u)} \\
&\approx \frac{P(\ell|a, u)P(a|t, u)P(t|u)}{\sum_{a'} P(\ell|a', u)P(a'|t, u)P(t|u)} \\
&= \frac{P(\ell|a, u)P(a|t, u)}{\sum_{a'} P(\ell|a', u)P(a'|t, u)}
\end{aligned} \tag{1}$$

In Equation 1, there are two major components, $P(\ell|a, u)$ and $P(a|t, u)$, which represent the location-activity correlation and time-activity correlation, respectively.

3.3 Mobility Inference, Deriving $P(\ell|a, t, u)$

Similar to Section 3.2, given u , a and t , the mobility inference, $P(\ell|a, t, u)$, can be simply as following steps:

$$\begin{aligned}
 P(\ell|a, t, u) &= \frac{P(\ell, a, t|u)}{P(a, t|u)} \\
 &\approx \frac{P(\ell|a, u)P(a|t, u)P(t|u)}{\int_{\ell'} P(\ell'|a, u)P(a|t, u)P(t|u)d\ell'} \\
 &= \frac{P(\ell|a, u)}{\int_{\ell'} P(\ell'|a, u)d\ell'} = P(\ell|a, u)
 \end{aligned} \tag{2}$$

From Equation 2, if we aim to infer individual mobility given user, activity and time factor, we could only take the activity factor into account for the approximation.

3.4 Time-Activity Model, $P(a|t, u)$

To evaluate time-activity model, $P(a|t, u)$, we observe the histogram of different activities over the time in a day in Figure 3.2. Figure 3.2 shows that the user has significantly different frequent patterns in one day under different activities. In [12], the authors exploit PLSA to infer spatial-temporal topics in social networks. In our work, we tackle only temporal activity inference in LBSNs. We think that the relations between activity and time are similar to the idea of collaborative filtering. The activity and time factor can be represented as different latent features. For example, the transportation activity is related to whether there is available train from office to home, and

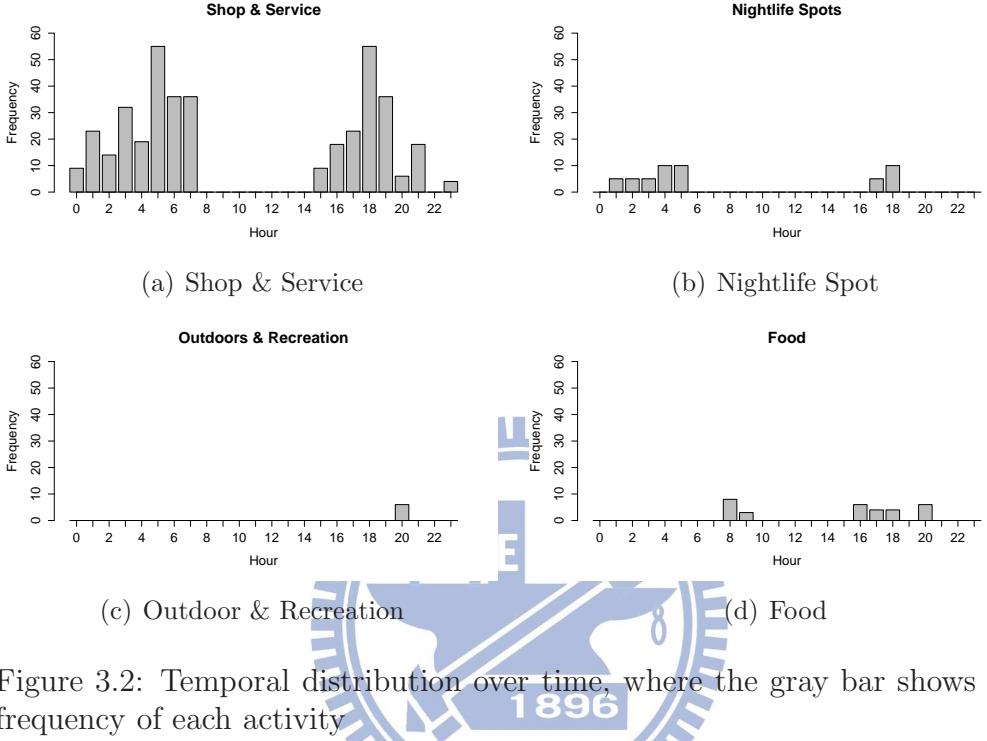


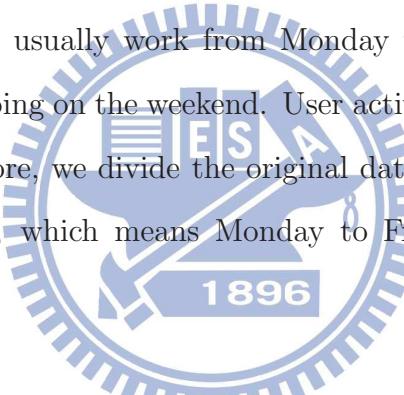
Figure 3.2: Temporal distribution over time, where the gray bar shows the frequency of each activity

there are some available trains from office to home in the afternoon. Thus, the probability is higher when the activity is transportation in the afternoon. To solve this problem, the intuitive solution is non-negative matrix factorization. However, non-negative matrix factorization is not suitable to deal with the sparsity issue of check-in records. For example, suppose there are 4 activity types and six time slots. If a user doesn't have any check-in with activity 3 and 4, after performing the non-negative matrix factorization, the predictive probability of the user doing activity 3 and 4 might be higher than doing activity 1 and 2. However, the probability of the user doing activity 3 and 4 should be lower than doing activity 1 and 2, due to the zero record of check-ins of activity label 3 and 4.

According to the above reson and idea, we propose a transition based

technique to capture the latent features between activity and time factor. First, we calculate the frequency of the activity with respect to the different time slots. As we mentioned before, user activity may be effected by the previous one. Based on those frequency in these time slots, we build an Activity Transition Model to find out the transition from an activity to the other activity. From this activity transition model, we calculate the probability for each activity in different time slots.

In addition, people usually work from Monday to Friday and do some entertainment or shopping on the weekend. User activity differs on weekdays and weekends. Therefore, we divide the original data into two parts, one is collected on weekdays, which means Monday to Friday, and the other is collected on weekends.



3.4.1 Activity Transition Model

The Activity Transition Mode captures the transition between each activity with respect to the previous one activity. The model is presented as below:

$$M(a_p, a_f) = \frac{\sum_{i=1:24n} F(a_p, s_{i-1}) \times F(a_f, s_i)}{\sum_{a_f \in ACT} \sum_{i=1:24n} F(a_p, s_{i-1}) \times F(a_f, s_i)} \quad (3)$$

where $M(a_p, a_f)$ represents the probability of activity a_f from activity a_p . $F(a_f, s_{i-1})$ shows the original frequency of activity a_f in the $i - 1$ th time slot s_{i-1} , and n is the number of hourly time slots, that is, $|s| = n$.

With different size of time slot length and different number of time slots, some blanks will appear. For example, a user has check-in records at 12:23, 12:25, 13:22, and 13:35. If the time slot lenght is an hour for this period, the

first two check-in records will be in slot 1, and the third and fourth records will be in slot 2. On the other hand, if there are four slots with length 30 minutes for each, the slot 2 will have no record. To fill in those empty slots, we calculate the probability by the Activity Transition Model for those empty slots. The equation is as below:

$$C(a_f, s_i) = \frac{F(a_f, s_i) + \sum_{a_p \in ACT} F(a_p, s_{i-1}) \times M(a_p, a_f)}{\sum_{a_f \in ACT} (F(a_f, s_i) + \sum_{a_p \in ACT} F(a_p, s_{i-1}) \times M(a_p, a_f))} \quad (4)$$

where the $C(a_f, s_i)$ represents the probability of activity a_f in the i th time slot s_i . $F(a_f, s_{i-1})$ shows the original frequent of activity a_f in the $i - 1$ th time slot s_{i-1} . $M(a_p, a_f)$ is the activity transition distribution of activity a_f from activity a_p . The Time-Activity Model is formed as following:

$$P(a|u, t) = C(a, t|u) \quad (5)$$

3.5 Location-Activity Model, $P(\ell|a, u)$

To capture the individual mobility with different activities, we analyze the check-in records with different activities. The results reflect users have different check-in patterns with different activities. Prior works utilize two-dimensional Gaussian distribution to model mobility from trajectories [1][5][2][4][11]. We summarize the idea of above works, we exploit Gaussian mixture model to capture individual mobility under different activities. The location-activity model $P(\ell|a, u)$ can be represented as follows:

$$P(\ell|a, u) = \sum_{i=1}^k \alpha_i \mathcal{N}(\mu_i, \Sigma_i) \quad (6)$$

where μ_i and Σ_i denote the mean and covariance matrix of i -th Gaussian distribution, respectively. k denotes the number of the Gaussian distributions in the model, and α_i denotes the probability of ℓ belonging to i -th Gaussian distribution.

To discover suitable parameters of the location-activity model for each user, we distinct check-in records into different activities, and exploit Expectation–maximization (EM) algorithm to learn the parameters of α_i , μ_i and Σ_i for each activity. To determine the best k for each activity, we run EM algorithm several times for k from 1 to 10, and select the k which has the greatest log-likelihood value.

3.6 Performing Activity and Mobility Inference

3.6.1 Constructing Time-Activity Model

To construct the Time-Activity Model, we first input the set of time t and activity a of a user u as our training set $TRAIN = (u, t, a)$, and define the initial value of two matrixes, which are *activity – time – weekday* and *activity – time – weekend*. Both of the matrixes are $|ACT| * |Slot|$ matrix, where $|ACT|$ is the number of activity types, $|Slot|$ is the number of slots, and the initial value are zeros. Second, we record the number of check-ins into

the corresponding matrix F . For example, a user u checks in at time t with activity label a on Friday. We add this check-in to $F[a, slot] = F[a, slot] + 1$. Then, construct the $|ACT| * |ACT|$ activity transition matrix M with Equation 3, where $M(a_p, a_f)$ is the activity transition matrix, which represents the transition probability of activity a_f from activity a_p , and $F(a_f, s_{i1})$ is the frequent matrix, which shows the original frequency of activity a_f in the $i - 1$ th time slot. Finally, fill both the *activity-time-weekday* and *activity-time-weekend* matrixes with Equation 4. Since different time slot length will lead to various results. We test the time slot length from 1 to 4 hours and choose the best length for each user as the output time-activity model.



3.6.2 Constructing Location-Activity Model

With input training set $TRAIN = (u, l, a)$, we construct the location-activity model with Equation 6. To fit the parameters of the location-activity model, we perform the Expectation-Maximization Algorithm (EM) to fit the parameters. Note that there may be more than one frequent region in the location distribution for each activity. Given k , we first label each location point randomly as $cluster1$, $cluster2$, ..., to $clusterk$. The parameters, namely μ and σ , of the first iteration are fitted by using maximum likelihood estimation. This fitting step is known as the *E-step*. According to the current μ and σ , we reassign these location points to the k clusters. This step is known as the *M-step*. This process is repeated until convergence. Since the EM algorithm is known to converge only to the local optima, we over-

come this problem by testing k from 1 to 10 and performing various initial region assignments to find the best fit with the highest likelihood.

3.6.3 Performing Activity Inference

After constructing both the time-activity model and the location-activity model, the entire model is shown as bellow:

```
{ "user1":{ "time-model":{ act-time-weekday, act-time-weekend }, "location-model":{ category1, category2, ... , category10 } }, ... "userN":{ ... } }
```

The model contains N subsets. Each subset stands for one user. For each user, there are *time – model* and *location – model*, which are constructed by *activity – time – model* and the *location – activity – model*, respectively. In *time – model*, there are two models for user doing activities on weekday and weekend, respectively. In *location – model*, there are 10 GMMs for activity/category from 1 to 10, respectively.

With input $TESTING = (u, t, \ell)$, where time t happened on weekday/weekend and is mapped to slot s , we calculate the probability for 10 categories $P(a_i) = act-time-weekday[a_i, slot] \times location-model(category_i, \ell)$ for $i = 1 : 10$. Then, choose the activity, which has the highest probability as our output.

3.6.4 Performing Mobility Inference

According to the input $TESTING = (u, t, a)$, where user u is doing activity a at time t , we can calculate the probability of each location ℓ . Then, select the location, which has the highest probability as our output.

Chapter 4

Performance Evaluation

4.1 Datasets Description

In this paper, we select the dataset, GeoText, for evaluating our proposed methods. The detail descriptions of the dataset are as follows:

GeoText¹: The GeoText dataset² is collected from Twitter over one week in March 2010 [3]. There are 9,475 users and 377,616 messages which have geographic information, and one user has 17 to 301 messages. Since we do not know the actual activity of each message, we use the location category instead of the activity. We use the location coordinate of each message to search the corresponding category via Foursquare API³, and select the first level category name as the activity. There are ten categories in the first level categories, Arts & Entertainment, Colleges & Universities, Events, Food, Nightlife Spots, Outdoors & Recreation, Professional & Other

¹The dataset is also adopted by [19].

²<http://www.ark.cs.cmu.edu/GeoText/>

³<https://api.foursquare.com/v2/venues/search>

Table 1: Statistics of two datasets

	GeoText
# of total users	9,475
# of check-ins	377,616
# of distinct locations	46,320
Max # of check-ins of a user	301
Min # of check-ins of a user	17
Period	March 2010

Table 2: The proportion of each activity in GeoText Dataset

Activity Type	GeoText
Arts & Entertainment (AE)	15,892
Colleges & Universities (CU)	16,697
Events (E)	106
Food (F)	46,332
Nightlife Spots (NS)	16,207
Outdoors & Recreation (OR)	39,832
Professional & Other Places (PO)	110,108
Residences (R)	24,387
Shops & Services (SS)	86,526
Travel & Transport (TT)	21,520

Places, Residences, Shops & Services and Travel & Transport. Table 1 shows the statistics in details, and Table 2 shows the activity distribution in the GeoText dataset.

4.2 Impact of Time Slot Length

In this experiment, we test different length of time slots for the temporal feature. Figure 4.1(a) shows the experiment result on GeoText dataset. The x-axis of Figure 4.1(a) shows the number of training size from 20 check-ins to 100 check-ins, and the y-axis represents the average settings of time slot length and the standard deviation of those users whos training check-ins are

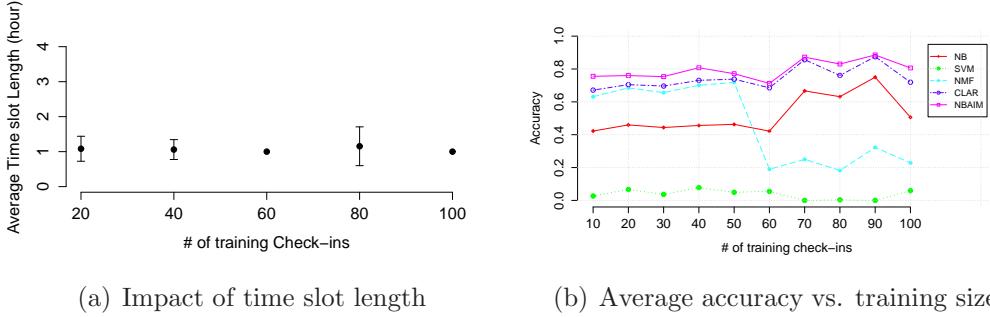


Figure 4.1: Impact of time slot length and the training size

20, 40, 60, 80, and 100, respectively. It shows that the best setting of time slot length for most users is an hour despite the training size.

4.3 Comparison of Activity Inference Methods

In this section, we will compare our Network Based Activity Inference Model(NBAIM) with the state-of-the-art approaches. First, we introduce the concepts of our baseline methods. Then, we will show the experiment results.

Baseline Methods We have four baseline methods to evaluate our proposed approach, namely Naive Bayesian(NB), support vector machine(SVM), non-negative matrix factorization(NMF), and collaborative location and activity recommendations[20](CLAR).

Figure 4.1(b) compares NBAIM with the baseline methods by calculating the average accuracy of users who have 10 to 100 training check-ins, respectively. Each line represents a baseline approach. From this experiment, we can see that even when there are only 10 training check-ins, NBAIM have

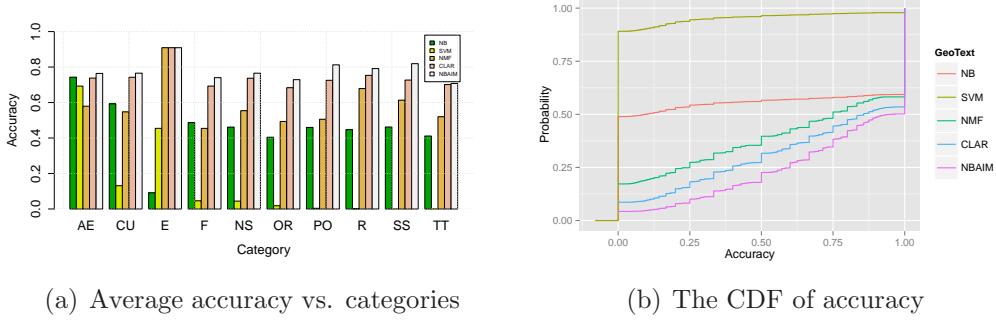


Figure 4.2: Comparison of activity inference methods

almost 80% accuracy, outperforming the other approaches. That is, even though the training data size of a user is very small, NBAIM can at least 70% precisely infer user activities.

Figure 4.2(a) shows the average accuracy on different categories. The x-axis of Figure 4.2(a) represents the ten categories, and the y-axis shows the average accuracy. Each bar with different colors represent different methods. It shows that NBAIM approach outperform the other methods for each category. For each category, the predictive accuracy are 70% or more.

Figure 4.2(b) shows the cumulative distribution function of accuracy. It shows that there are 50% or more users have 88% or more accuracy.

4.4 Comparison of Location Inference Methods

Baseline Method We compare NBAIM with the baseline method which is kernel density estimation(KDE). That is, we utilize the kernel density estimation in the location-activity model, instead of using Gaussian Mixture Model.

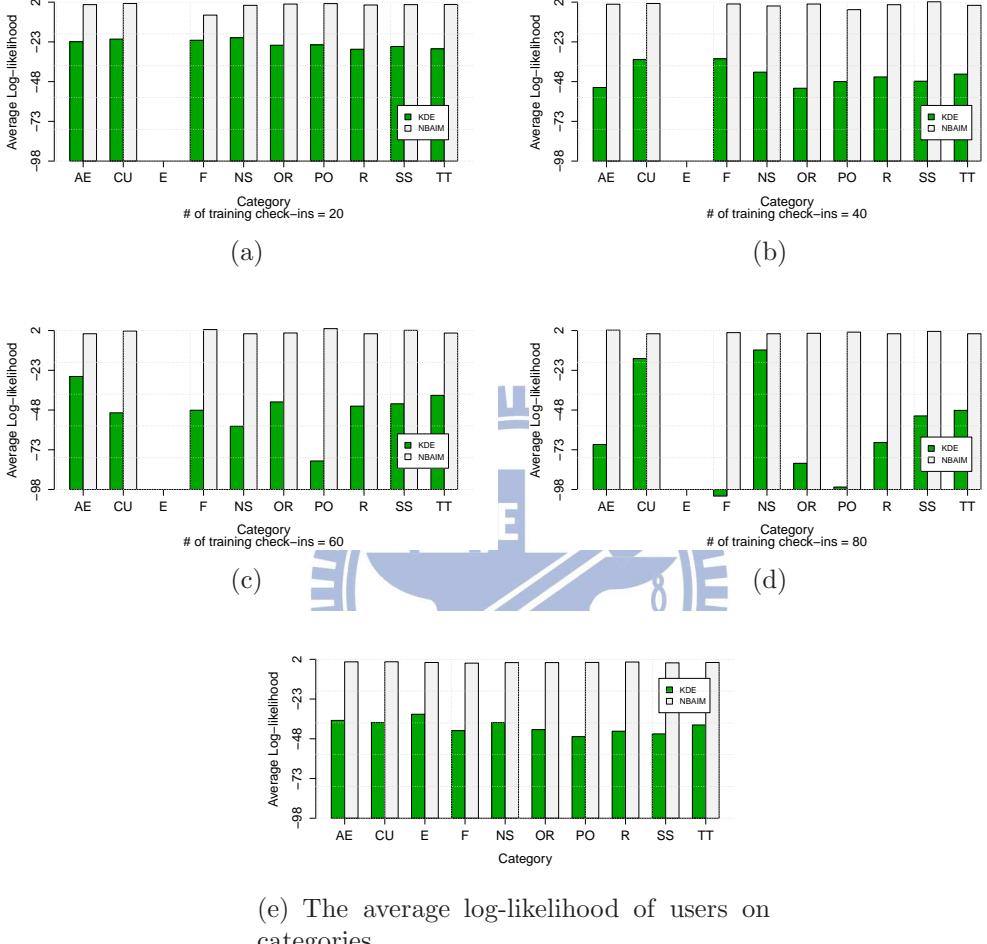


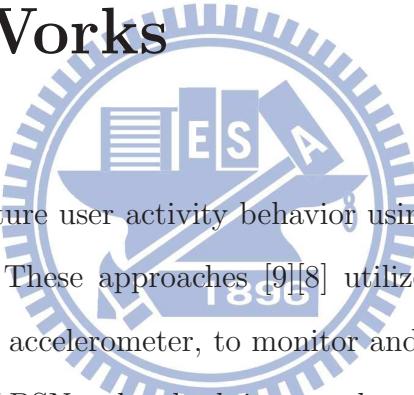
Figure 4.3: Comparison of location inference methods

Similarly, we build a KDE model for each users and each activities. Besides, we compare the performance by calculating the average log-likelihood.

Figure 4.3 shows the average log-likelihood on ten categories when the training size are 20, 40, 60, and 80, respectively. The two bars with different colors represent different methods. From Figure 4.3 and Figure 4.3(e), we can see that despite the training data size of each user, NBAIM outperforms KDE on all of the categories.

Chapter 5

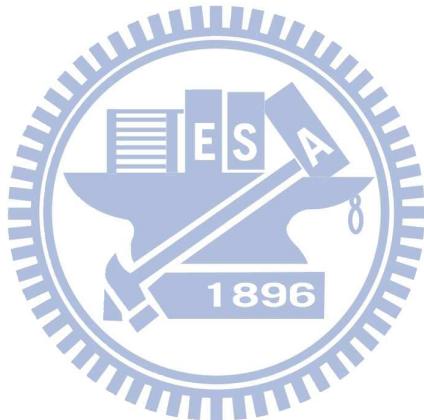
Related Works



Prior works try to capture user activity behavior using sensors equipped on mobile devices [9][8]. These approaches [9][8] utilize the sensors, such as location-aware sensors, accelerometer, to monitor and recognize the activity of users. However, in LBSNs, the check-in records only contain spatial and temporal information. Moreover, the check-in records have spatial-temporal sparsity issue. It means that a user may only have less than 100 check-in records. Thus, the approaches using sensors [9][8] are not eligible to infer individual user activity in LBSNs.

Some approaches focus on infer user activity from trajectories [21][20][10][14]. In [14][21], the authors try to detect the transportation modes from trajectories, such as walking and driving. In [20], the authors discover the relations of activity and location, and exploit the matrix factorization technique to model the relations. However, the model can not extract personalized location activity. Moreover, the environment of LBSNs does not allow the platforms sustained collect data. Thus, the above approaches are not eligible in LBSNs.

In LBSNs, [17], the authors try to guess the tags of each location in LBSNs based on check-in records. However, the tags describe each location for all users. Some locations contain multiple facilities. For example, a university has offices, classrooms and food court. Users visit the university have different activities, such as studying and dining. Therefore, the above approve approaches [21][20][17] are not eligible to solve our target problem.



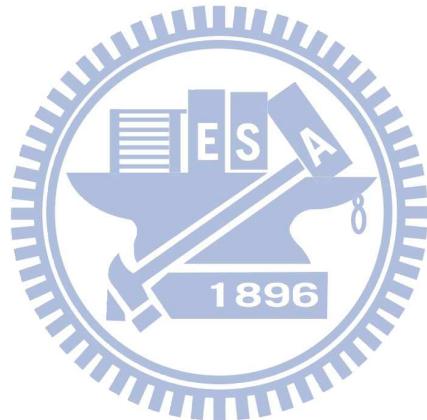
Chapter 6

Conclusion

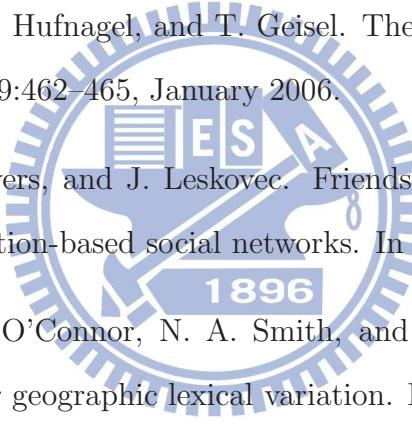


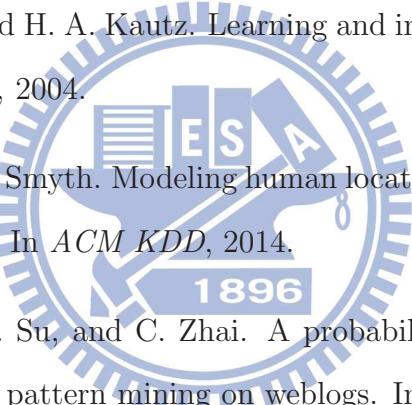
In this paper, we formulate the individual activity and mobility inference problem for personalized location-based service in location based social networks. We proposed the Bayesian based Activity Inference Model for user activity and mobility inference. We utilize Bayesian network to describe the relations among spatial, temporal and activity factors. There are two component in our Network Based Activity Inference Model, which are time-activity model and location-activity model. Instead of utilizing the traditional approach, which is non-negative matrix factorization, we proposed the activity transition model to closely capture the transition between user activities. On the other hand, we utilize the Gaussian Mixture Model to model user mobility, and compare with the Kernel Density estimation. We experiment on a real dataset called GeoText. Our personalized model outperform the state-of-art methods, which are NB, SVM, NMF, and CLAR for activity inference, and outperforms the KDE approach for mobility inference. With the experiment results, it shows that our activity inference model have average 70%

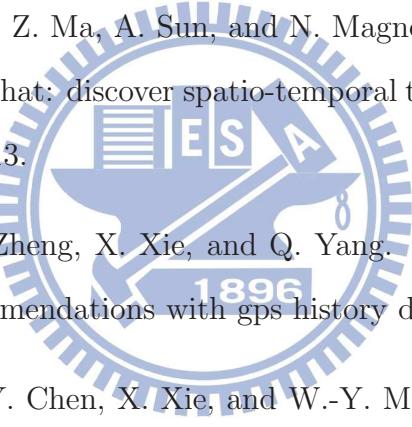
or more. Also the average log-likelihood of our mobility inference model is nearly 1. Therefore, NBAIM has good performances on both activity and mobility inference.



Bibliography

- 
- [1] D. Brockmann, L. Hufnagel, and T. Geisel. The scaling laws of human travel. *Nature*, 439:462–465, January 2006.
 - [2] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *ACM KDD*, 2011.
 - [3] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing. A latent variable model for geographic lexical variation. In *EMNLP*, 2010.
 - [4] H. Gao, J. Tang, X. Hu, and H. Liu. Modeling temporal effects of human mobile behavior on location-based social networks. In *ACM CIKM*, 2013.
 - [5] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453:779–782, June 2008.
 - [6] J. Han, M. Kamber, and J. Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 3rd edition, 2011.
 - [7] L. Hong, A. Ahmed, S. Gurumurthy, A. J. Smola, and K. Tsoutsouliakis. Discovering geographical topics in the twitter stream. In *WWW*, 2012.

- 
- [8] N. D. Lane, L. Pengyu, L. Zhou, and F. Zhao. Connecting personal-scale sensing and networked community behavior to infer human activities. In *ACM UbiComp*, 2014.
 - [9] N. D. Lane, Y. Xu, H. Lu, S. Hu, T. Choudhury, A. T. Campbell, and F. Zhao. Enabling large-scale human activity inference on smartphones using community similarity networks (CSN). In *ACM UbiComp*, 2011.
 - [10] L. Liao, D. Fox, and H. A. Kautz. Learning and inferring transportation routines. In *AAAI*, 2004.
 - [11] M. Lichman and P. Smyth. Modeling human location data with mixtures of kernel densities. In *ACM KDD*, 2014.
 - [12] Q. Mei, C. Liu, H. Su, and C. Zhai. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *WWW*, 2006.
 - [13] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil. An empirical study of geographic user activity patterns in foursquare. In *AAAI ICWSM*, 2011.
 - [14] D. J. Patterson, L. Liao, D. Fox, and H. A. Kautz. Inferring high-level behavior from low-level sensors. In *UbiComp*, 2003.
 - [15] D. Peebles, H. Lu, N. D. Lane, T. Choudhury, and A. T. Campbell. Community-guided learning: Exploiting mobile sensor users to model human behavior. In *AAAI*, 2010.
 - [16] S. Sizov. Geofolk: Latent spatial semantics in web 2.0 social media. In *ACM WSDM*, 2010.

- 
- [17] M. Ye, D. Shou, W.-C. Lee, P. Yin, and K. Janowicz. On the semantic annotation of places in location-based social networks. In *ACM KDD*, 2011.
 - [18] M. Ye, P. Yin, W.-C. Lee, and D.-L. Lee. Exploiting geographical influence for collaborative point-of-interest recommendation. In *ACM SIGIR*, 2011.
 - [19] Q. Yuan, G. Cong, Z. Ma, A. Sun, and N. Magnenat-Thalmann. Who, where, when and what: discover spatio-temporal topics for twitter users. In *ACM KDD*, 2013.
 - [20] V. W. Zheng, Y. Zheng, X. Xie, and Q. Yang. Collaborative location and activity recommendations with gps history data. In *WWW*, 2010.
 - [21] Y. Zheng, Q. Li, Y. Chen, X. Xie, and W.-Y. Ma. Understanding mobility based on gps data. In *UbiComp*, 2008.
 - [22] Z. Zhu, U. Blanke, A. Calatroni, and G. Tröster. Human activity recognition using social media data. In *MUM*, 2013.