Who, Where, When and What: Discover Spatio-Temporal Topics for Twitter Users

Quan Yuan, Gao Cong, Zongyang Ma, Aixin Sun, Nadia Magnenat-Thalmann
School of Computer Engineering, Nanyang Technological University, Singapore 639798
{qyuan1@e., gaocong@, zma4@e., axsun@, nadiathalmann@}ntu.edu.sg

ABSTRACT

Micro-blogging services, such as Twitter, and location-based social network applications have generated short text messages associated with geographic information, posting time, and user ids. The availability of such data received from users offers a good opportunity to study the user's spatial-temporal behavior and preference. In this paper, we propose a probabilistic model W^4 (short for **W**ho+**W**here+**W**hen+**W**hat) to exploit such data to discover individual users' mobility behaviors from spatial, temporal and activity aspects. To the best of our knowledge, our work offers the first solution to jointly model individual user's mobility behavior from the three aspects. Our model has a variety of applications, such as user profiling and location prediction; it can be employed to answer questions such as "Can we infer the location of a user given a tweet posted by the user and the posting time?" Experimental results on two real-world datasets show that the proposed model is effective in discovering users' spatial-temporal topics, and outperforms state-of-the-art baselines significantly for the task of location prediction for tweets.

Categories and Subject Descriptors

H.2.8 [**Database applications**]: Data mining; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

Keywords

User Profiling; Graphical Model; Prediction and Recommendation; Spatio-Temporal; Twitter

1. INTRODUCTION

Posting short messages through micro-blogging services (*e.g.,* Twitter and Tumblr) has become an indispensable part of the daily life for many users. For example, as of December 2012, there were more than 200 million monthly active Twitter users[1]. A short text message posted through Twitter is known as a *tweet* with the maximum length of 140 characters. With the prevalence of GPS-enabled devices (*e.g.,* smart phones), many tweets are associated with location information. The locations may be in the form of the latitude and longitude coordinates, or in the form of exact addresses. The latter can be specified explicitly by users, detected by mobile devices, or geo-tagged by geo-tagging tools. Apart from Micro-blogging services, Location-based Social Network (LBSN) applications, such as Foursquare and Facebook Places, allow users to share their current locations and activities by checking-in points of interests and composing short text messages at check-ins. In short, each the geo-annotated tweet or check-in message contains a user id, a text message, posting time, and a location.

The large availability of such short messages directly received from users offers an exciting opportunity to study the behaviors of individuals (***who*** is the user?) with respect to three important aspects, namely, geographical location (***where*** does an individual visit?), time (***when*** does a user visit a place for some activity?), and activity (***what*** does a user do?).

To the best of our knowledge, most of previous studies on modeling mobility behaviors of individual users have focused on at most three out of the four factors. Several studies [4, 5, 8, 11, 22] focus on the geographical location and the temporal factors, aiming at modeling and analyzing the relationship between user's mobility patterns and temporal factor. An example finding of such studies would be that a user usually visits a region centered at a particular building at 2-3pm. Note that understanding human mobility has many applications, such as location-based recommendation and location-based advertisement among others. However, these studies ignore the activity aspects of users, which is represented as tweet content. There also exist studies [13] that focus on the geographic location and activity aspects, but ignore the temporal factor. An example finding captured by the models in these studies is that a user participates in some activities (*e.g.,* having meals) in a geographic region. However, these models cannot capture the relationship with the temporal factor.

In this paper, we model the interactions of all the four factors in a unified way to better understand individuals' behaviors. In particular, we discover spatial-temporal topics and identify uses' interest over time and regions from the geo-annotated messages (*e.g.,* tweets) from users. With our model, we are able to answer the following questions among others.

- Can we predict the activity of a user at a given time?
- Can we infer the location of a user given a tweet posted by the user
- What are the mobility pattern and words used by a user at a given time?
- Can we infer the user who will visit a given location at a given time?

[1] `https://twitter.com/twitter/status/281051652235087872`. Accessed 20 Feb 2013

KDD'13, August 11–14, 2013, Chicago, Illinois, USA.

It is however challenging to develop a model to capture the four factors jointly, since they are in different data types (continuous and discrete), and the four dimensions together will make the modeling and parameter estimation complicated. Moreover, the interdependencies among them and role played by each is unclear. To this end, based on several intuitions (to be detailed in Section 3.1), we propose a novel probabilistic generative model to model user behavior from the geographic, temporal, and activity aspects, which has a variety of applications such as user profiling, content recommendation, location prediction and recommendation, topic tracking, etc. We show that our model is able to identify interesting spatial-temporal topics for users, and we demonstrate its effectiveness on various applications such as predicting locations for tweets (*w/wo* time), predicting locations for users at a given time, and predicting users who will visit a given location at a given time. Experimental results show that our approach outperforms the existing approaches [8, 13, 17] for these applications.

The contributions of this work are summarized as follows:

1. We propose a novel probabilistic model W^4, which is short for **W**ho+**W**here+**W**hen+**W**hat, to model users' mobility behaviors from geographic, temporal and activity aspects in a unified way. The model enables us to discover geographical-temporal topics for individual users.
2. We propose new inference algorithms for estimating the model parameters.
3. Experimental results on two real-world datasets demonstrate that our model is capable of identifying interesting spatial-temporal topics for users. The results also show that our model outperforms the state-of-the-art methods significantly for various applications including location prediction and user prediction.

The rest of this paper is organized as follows: We survey the related work in Section 2. Section 3 presents the proposed model and the method of estimating model parameters. We discuss some applications of our model in Section 4, and present the experimental results in Section 5. Section 6 concludes our work.

2. RELATED WORK

We group the existing proposals on mobility modeling and geographical topic modeling based on the aspects considered in these proposals, namely Who, Where, When and What.

Where What: The existing studies on geographical topic modeling focus on the geographic (Where) and activity (What) aspects, but do not consider users at all. How to represent locations is an essential part of these studies. Locations have two properties: the geo-locations represented by coordinates, and the functions (e.g., a shop) represented by the topics. Based on the ways of representing locations, the existing studies can be divided into two categories:

First, some proposals [12,23] represent locations by location ids, and this enables these proposals to distinguish the functions between locations. However, this modeling manner fails to exploit the coordinate information, which is important to analyze the user mobility region. Specifically, Wang et al. [23] propose a Latent Dirichlet Allocation (LDA) based model to learn the relationship between location and words. They assume that each word is associated with a location. When a word is generated, its associated location is also generated. Hao et al. [12] mine the location-representative topic from travelogues using an LDA-based model.

Second, other proposals [9, 21, 26] represent locations as coordinates, and they are capable of describing the mobility regions of users. However, they either neglect the functions of locations or assume that nearby locations have the same functions, which are generally not true in reality. Eisentein et al. [9] propose regional variants of topics, which are used to generate the words of a geo-referenced document. They use bi-variant Gaussian distributions of regions to generate coordinates of locations. Sizov [21] proposes GeoFolk model to manage geo-referenced documents. In addition to the word distribution, each topic in GeoFolk is also associated with two Gaussian distributions over latitude and longitude, respectively. In GeoFolk each geographic region represents a distinct topic/function. Hence, it fails to correlate the different regions with the same function; it would not be suitable to model a large area containing many topical regions since the topic model becomes computationally expensive as the number of topics is large. Yin et al. [26] propose a Probabilistic Latent Semantic Analysis (PLSA) based model to discover geographical topics. In the model, each region is characterized by a topic distribution, and represented by a bi-variant Gaussian distribution over coordinates.

In contrast, we propose an approach that is able to exploit both properties of locations. Further, different from these proposals, we model individual users and consider the temporal aspect.

Where When What: Mei et al. [18] model topics of documents from spatio-temporal aspects using PLSA. Specifically, they assume that each word is drawn from a background word distribution, a time and location dependent topic, or a topic of the documents. Similarly, Bauer *et al.* propose an LDA-based spatio-temporal model [1], where a city is divided into grids. Compared with the models [1, 18], our model considers more aspects:1) the models [1, 18] do not consider the user information at all; 2) it either does not consider the geographic property of locations [18], or does not consider the functions of locations [1]; 3) they only consider discretized time. There are also several works on extracting events from twitter stream [16, 19], which exploit the temporal (When) and activity (What) information, and some work even considers the geographic aspect (Where) [20]. However, their problem settings are different from ours, and none of them considers user information.

Who Where When: We next review the work on modeling mobility behaviors of individual users (Who) that focuses on the geographic (Where) and temporal (When) aspects.

Brockmann et al. [4] find that human mobility behavior can be approximated by the continuous-time random-walk model. Gonzĺélez et al. [11] find that users periodically return to a few previously visited locations, such as home or office, and the mobility of each user can be represented by a stochastic process centered at a fixed point. Song et al. [5,22] focus on the predictability in human mobility, and report that there is a 93% predictability of human mobility, which is contributed by the high regularity of human behavior. Cho et al. [8] observe that the mobility of each user is centered at two regions (representing "work" and "home"), and model each region as a Gaussian distribution over latitude and longitude. The probability that a user stays at the two regions is modeled as a function of time. They propose a generative model, Periodic Mobility Model (PMM), to predict the location of a user. PMM takes a user and time as input; It generates a region, and the region further generates a geo-location.

None of these studies consider the activity (topic) aspect of user behavior as we do in this paper.

Who Where What: The recent work [13] presents a model from the geographic (Where) and activity (What) aspects for individuals (Who). Hong et al. [13] propose a method to learn the geographical topics for twitter users. For a user, this method first generates a region based on the popularity of regions and the preference of the user over the regions. Then, a topic is generated dependent on both the region and the user. The topic, together with the region, gener-

ates the words of a tweet; the region alone generates the coordinates based on its Gaussian distribution over coordinates.

Different from our work, the work does not consider the temporal aspect. In addition, the regions [13] are global, which are shared by all users, and cannot precisely depict individual users' mobility areas, while our proposed model is able to model regions of individuals. Moreover, the method fails to consider the semantic information of individual venues in the same region.

In summary, none of existing studies aim to model the three aspects (Where, When, and What) for individual users (Who). In addition, previous work does not exploit the coordinates and functions of locations simultaneously, and thus they cannot capture both the geo-graphic region and functional information of locations.

3. PROPOSED MODEL

We present the proposed approach W^4 to modeling user mobility behavior with a collection of geo-tagged tweets. We first describe the intuitions of W^4 in Section 3.1 We then present the model in Section 3.2, and detail the inference algorithm in Section 3.3, followed by the complexity analysis of the algorithm in Section 3.4.

3.1 Intuitions

We model user mobility behavior based on the following intuitions. These intuitions jointly cover the four factors in user mobility behavior (*i.e.,* who, where, when, and what).

1. **Intuition 1**: an individual's mobility usually centers at different personal geographical regions, *e.g.,* home region and work region [8] and users tend to visit places within these regions. In addition, the region where a user stays is influenced by the time factor, *i.e.,* the time in a day and the day of a week. It has been reported that users often demonstrate different mobility patterns in weekdays and weekends [8]. For instance, most users have lunch at different places during weekdays and weekends, and a user is likely to stay at different regions in the noon and evening.
2. **Intuition 2**: the topics of a user at a place are influenced by both the user's personal topic preferences and the region where the user stays. For example, suppose a user who is interested in both eating and hiking comes to a place full of restaurants, the user is more likely to be interested in the eating topic. In addition, the topics of a user at her home region (*e.g.,* entertainment and shopping) are expected to be different from the work-related topics at her work region.
3. **Intuition 3**: when a user chooses a location to visit, both the topic requirement and the region where the user stays should be considered. Intuitively, a user tends to visit nearby locations within her current region of stay that meet her requirement (*e.g.,* for meal).
4. **Intuition 4**: different regions and different topics lead to different language variations, which in turn reflect the user's activity. Therefore, the words in user's tweet are affected by both the topic and the region. For example, if a user is shopping at her home region, the words she would use are related to both the shopping topic and home region, such as "grocery", "family", etc.

3.2 Notations and Model Description

We consider each user u has several personal regions, *i.e.,* home region and work region, denoted by $\{r_{u,0}, r_{u,1}, ..., r_{u,|R|}\}$, where $|R|$ is the number of regions. The personal regions are estimated based on the locations of all geo-tagged tweets from a user. We model a

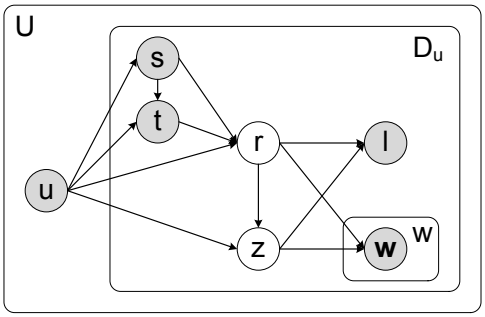


Figure 1: The graphical representation of proposed model W^4

location ℓ as a two-tuple $\ell = \{id_\ell, \boldsymbol{cd}_\ell\}$, where id_ℓ is the identifier of the location, and $\boldsymbol{cd}_\ell$ is the latitude and longitude coordinates of the location, denoted by $\boldsymbol{cd}_{\ell,0}$ and $\boldsymbol{cd}_{\ell,1}$, respectively. A region r is modeled by a bi-variant Gaussian over the latitude and longitude, parameterized by the mean vector $\boldsymbol{\mu}_r$ and covariance matrix $\boldsymbol{\Sigma}_r$. Note that we use r to represent a region (*i.e.,* any one of the personal regions) when the semantic is clear.

To model the time factor, we model time t in a day as a continuous variable in $\{hh : mm : ss\}$ format, and categorize days into two classes, namely, weekdays and weekends. Specifically, we use $s \in \{0, 1\}$ to denote a day of a week, *i.e.,* $s = 0$ for a weekday and $s = 1$ for a weekend day. Note that t is cyclical on a daily basis. For instance, the time difference between 23:00:00 and 1:00:00 is the same as the difference between 1:00:00 and 3:00:00.

With the above notations, we consider a tweet d is a five-tuple $d = \{u_d, \ell_d, \boldsymbol{w}_d, t_d, s_d\}$, where u_d denotes the user or the author of the tweet; ℓ_d, t_d, and s_d denote the location, the time in a day, and the day of a week, as described earlier; $\boldsymbol{w}_d$ are the words in tweet d. For easy presentation, we use D, U, and L to denote the collection of tweets, users, and locations respectively. The word vocabulary is denoted by V. That is, $d \in D$, $u \in U$, $\ell \in L$, and each word w_i in $\boldsymbol{w}_d$ belongs to V. The topics of a user are reflected by the words in the user's tweets.

Based on the aforementioned intuitions and notations, W^4 generates the day, time, words, and location for each tweet posted by a user, shown in Figure 1. The generative process is briefly described below. The details of the distributions are discussed after the generative process, followed by the inference algorithm in Section 3.3.

1. For each tweet d of a given user u, a day s is first selected based on a Bernoulli distribution $p(s|u)$, and then a time in that day t is selected based on $p(t|u, s)$, which can be a uniform distribution or Gaussian mixture distribution, among other appropriate distributions. After that, a personal region r is generated by drawing from the multinomial distribution $p(r|u, s, t)$ (**Intuition 1**).
2. Parameterized by the topic preference of the user u and the sampled region r, a topic z is generated using the multinomial distribution $p(z|u, r)$ (**Intuition 2**).
3. After generating the region and the topic, the location ℓ and each word w are sampled based on $p(w|r, z)$ and $p(\ell|r, z)$, respectively (**Intuition 3** and **4**).

In summary, to generate a collection of tweets D, the following generative process is applied to each user $u \in U$:

- For each tweet $d \in D_u$, D_u is the collection of tweets posted by u
  - Draw a day $s \sim p(s|u)$;

– Draw a time $t \sim p(t|u,s)$;
– Draw a region $r \sim p(r|u,s,t)$;
– Draw a topic $z \sim p(z|u,r)$;
– Draw a location $\ell \sim p(\ell|r,z)$;
– For each word w in $\boldsymbol{w}_{d_i}$, draw $w \sim p(w|r,z)$.

While the distributions for modeling $p(s|u)$ and $p(t|u,s)$ are relatively straightforward, it is complicated to model $p(r|u,s,t)$, given that r is discrete while t is continuous. We propose a method to solve this problem, which will be detailed in Section 3.3. To model $p(w|r,z)$, a parameter λ is introduced to balance the importance between the region and the topic, *i.e.,* $p(w|r,z) = \lambda p(w|z) + (1-\lambda)p(w|r)$, where $p(w|z)$ and $p(w|r)$ are the word distribution of topic z and region r, respectively. For sampling, an indicator x following a Bernoulli distribution is assumed with probability $p(x=0) = \lambda$ and $p(x=1) = 1-\lambda$. A word w is sampled based on $p(w|z)$ if $x=0$, and sampled based on $p(w|r)$ if $x=1$. Because a tweet is very short, we assume all words in a tweet come from the same topic.

Next, we generate a location according to r and z. It is however challenging to model the generating process of location. Previous studies treat a location either as geographic coordinates or a location identifier. Treating locations as geographic coordinates makes it feasible to capture user's mobility regions [26], or discover geo-regions that have specific topics [13, 21], but fails to capture the topic variations of different locations. On the other hand, treating locations as location identifiers enables us to differentiate the topics of locations [12, 18, 23] (because $p(\ell|z)$ is always modeled by a multinomial distribution, which calls for a limited location set L), but the geographic coordinate information is ignored.

As stated in **Intuition 3**, a user tends to visit a nearby location (*e.g.,* restaurant) that can fulfill her topical needs (*e.g.,* lunch). That is, when choosing a location to visit, a user jointly considers both its geographic location and its topic (*e.g.,* restaurant or bar). However, no previous work jointly models geographic locations and topics of locations. Indeed, it is hard to model them together: from the geographic perspective, a location is drawn from a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r)$, which is a continuous distribution, while from the topic perspective, a location is generated based on a multinomial distribution $p(\ell|z)$, which is discrete. What makes the issue even more complicated is that if we treat $p(\ell|r)$ as the density of its Gaussian distribution at ℓ, $p(\ell|r)$ will have a different scale from that of $p(\ell|z)$. The former will be much greater than 1 if ℓ is close to the mean vector $\boldsymbol{\mu}_r$. Thus, a simple combination by linear interpolation of the two components leads to $p(\ell|z)$ overwhelmed by $p(\ell|r)$. A straightforward solution is to perform a good number of sampling based on the Gaussian distribution, and estimate $p(\ell|r)$ by counting the times each location is sampled. However, this approach will introduce inner loops, which will greatly deteriorate the efficiency. To solve this problem, we propose a method to compute the probability of generating ℓ, given a region r, according to Lemma 1:

$$p(\ell|r) \propto \frac{exp(-\frac{1}{2}(\widetilde{cd}_{\ell,0}^2 + \widetilde{cd}_{\ell,1}^2)) - exp(-\frac{1}{2}(\widetilde{cd}_{\ell',0}^2 + \widetilde{cd}_{\ell',1}^2))}{\pi(\widetilde{cd}_{\ell',0}^2 + \widetilde{cd}_{\ell',1}^2 - \widetilde{cd}_{\ell,0}^2 - \widetilde{cd}_{\ell,1}^2)},$$

where $\widetilde{cd}_\ell$ and $\widetilde{cd}_{\ell'}$ are the geographic coordinates of ℓ and its close point ℓ' after we perform the standardized coordinate transformation for Gaussian as follows. For each point $\boldsymbol{cd} \sim f(\boldsymbol{cd}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{2\pi|\boldsymbol{\Sigma}|^{\frac{1}{2}}} exp(-\frac{1}{2}(\boldsymbol{cd}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{cd}-\boldsymbol{\mu}))$, we replace it with $\widetilde{\boldsymbol{cd}} = \boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{cd}-\boldsymbol{\mu})$, then we have $\widetilde{\boldsymbol{cd}} \sim \frac{1}{2\pi} exp(-\frac{\widetilde{cd}_0^2+\widetilde{cd}_1^2}{2})$ (Figure 2 (b)).

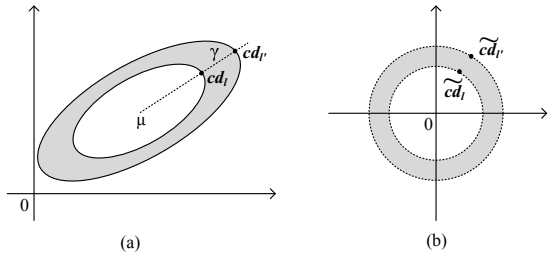


Figure 2: Standardizing Gaussian to calculate $p(\ell|r)$

Lemma 1:

$$p(\ell|r) \propto \frac{exp(-\frac{1}{2}(\widetilde{cd}_{\ell,0}^2 + \widetilde{cd}_{\ell,1}^2)) - exp(-\frac{1}{2}(\widetilde{cd}_{\ell',0}^2 + \widetilde{cd}_{\ell',1}^2))}{\pi(\widetilde{cd}_{\ell',0}^2 + \widetilde{cd}_{\ell',1}^2 - \widetilde{cd}_{\ell,0}^2 - \widetilde{cd}_{\ell,1}^2)}.$$

Proof: Draw a line from the mean point of a Gaussian $f(\boldsymbol{cd}|\boldsymbol{\mu},\boldsymbol{\Sigma})$ to a point $\boldsymbol{cd}_\ell$, and on the line we can get another point $\boldsymbol{cd}_{\ell'}$ that is γ farther from the mean point than $\boldsymbol{cd}_\ell$, where γ is a small value (Figure 2 (a)). If γ is small enough, we can assume that the points between the contours defined by $\boldsymbol{cd}_\ell$ and $\boldsymbol{cd}_{\ell'}$ have equal probability. After transformation, the original $\boldsymbol{cd}_\ell$ and $\boldsymbol{cd}_{\ell'}$ become $\widetilde{\boldsymbol{cd}}_l$ and $\widetilde{\boldsymbol{cd}}_{\ell'}$, which define an annulus a with area $\pi(\widetilde{cd}_{\ell',0}^2 + \widetilde{cd}_{\ell',1}^2) - \pi(\widetilde{cd}_{\ell,0}^2 + \widetilde{cd}_{\ell,1}^2)$. By integral in the polar coordinate system, we get the probability of a as follows.

$$\begin{aligned} p(a) &= (1 - exp(-\frac{1}{2}(\widetilde{cd}_{\ell',0}^2 + \widetilde{cd}_{\ell',1}^2))) - (1 - exp(-\frac{1}{2}(\widetilde{cd}_{\ell,0}^2 + \widetilde{cd}_{\ell,1}^2))) \\ &= exp(-\frac{1}{2}(\widetilde{cd}_{\ell,0}^2 + \widetilde{cd}_{\ell,1}^2)) - exp(-\frac{1}{2}(\widetilde{cd}_{\ell',0}^2 + \widetilde{cd}_{\ell',1}^2)). \end{aligned}$$

Dividing $p(a)$ by its area, we get the probability value

$$\frac{exp(-\frac{1}{2}(\widetilde{cd}_{\ell,0}^2 + \widetilde{cd}_{\ell,1}^2)) - exp(-\frac{1}{2}(\widetilde{cd}_{\ell',0}^2 + \widetilde{cd}_{\ell',1}^2))}{\pi(\widetilde{cd}_{\ell',0}^2 + \widetilde{cd}_{\ell',1}^2 - \widetilde{cd}_{\ell,0}^2 - \widetilde{cd}_{\ell,1}^2)}.$$

After the transformation, we get a multinomial distribution $p(\ell|r)$, which has the same scale with $p(\ell|z)$. Then, a parameter κ is introduced to balance $p(\ell|z)$ and $p(\ell|r)$, *i.e.,* $p(\ell|r,z) = \kappa p(\ell|z) + (1-\kappa)p(\ell|r)$. □

3.3 Inference Algorithm

As shown in Figure 1, there are two latent variables in W^4, namely, region r and topic z. The joint probability over tweet $d = \{u_d, \ell_d, \boldsymbol{w}_d, t_d, s_d\}$, region r, and topic z, can be written as:

$$\begin{aligned} p(d,r,z) =& p(u_d, r, z, s_d, t_d, \ell_d, \boldsymbol{w}_d) \\ =& p(u_d)p(s_d|u_d)p(t_d|u_d,s_d)p(r|u_d,s_d,t_d) \\ & p(z|u_d,r)p(\ell_d|r,z)p(\boldsymbol{w}_d|r,z), \end{aligned} \quad (1)$$

where

$$\begin{aligned} p(\ell_d|r,z) =& \kappa p(\ell_d|z) + (1-\kappa)p(\ell_d|r), \\ p(\boldsymbol{w}_d|r,z) =& \prod_{w \in \boldsymbol{w}_d} (\lambda p(w|z) + (1-\lambda)p(w|r))^{c(w,\boldsymbol{w}_d)}. \end{aligned}$$

In the above equation, $c(w, \boldsymbol{w}_d)$ is the count of word w in $\boldsymbol{w}_d$.

We use an indirect way to calculate $p(r|u_d, s_d, t_d)$. Specifically, from Figure 1, we find the nodes u, s, t, r and the edges between them form a fully connected graph, and other nodes, namely, z, ℓ and $\boldsymbol{w}$ are all children of them. For this fully connected graph, we can re-order its nodes as follows [3]:

$$p(u)p(s|u)p(t|u,s)p(r|u,s,t) = p(u)p(s|u)p(r|u,s)p(t|u,s,r), \quad (2)$$

where $p(t|u,s,r)$ follows Gaussain distribution parameterized by the mean $\nu_{u,s,r}$ and variance $\sigma^2_{u,s,r}$, *i.e.,*, given a day s, the time a user u stay within a region r is centered at the time $\nu_{u,s,r}$, and the probability of staying at r decreases as the time becomes derivative from $\nu_{u,s,r}$.

Substituting Equation 2 into Equation 1, we have a new expression of the joint probability:

$$\begin{aligned} p(d,r,z) =& p(u_d)p(s_d|u_d)p(r|u_d,s_d)p(t_d|u_d,s_d,r) \\ & p(z|u_d,r)p(\ell_d|r,z)p(\boldsymbol{w}_d|r,z). \end{aligned} \tag{3}$$

We can also prove it in a different way: by applying Bayes Theorem to $p(r|u_d,s_d,t_d)$, we have:

$$\begin{aligned} p(r|u_d,s_d,t_d) =& \frac{p(u_d,r,s_d,t_d)}{p(u_d,s_d,t_d)} \\ =& \frac{p(t_d|u_d,s_d,r)p(u_d,s_d,r)}{p(u_d,s_d,t_d)} \\ =& \frac{p(t_d|u_d,s_d,r)p(u_d)p(s_d|u_d)p(r|u_d,s_d)}{p(u_d)p(s_d|u_d)p(t_d|u_d,s_d)}. \end{aligned} \tag{4}$$

By substituting Equation 4 into Equation 1, we again reach Equation 3.

This model has a set of parameters $p(r|u,s)$, $p(z|u,r)$, $\nu_{u,s,r}$, $\sigma_{u,s,r}$, $p(\ell|z)$, $\boldsymbol{\mu}_{u,s,r}$, $\boldsymbol{\Sigma}_{u,s,r}$, $p(w|z)$ and $p(w|r)$. Denoting them by Ψ, we have the log-likelihood of the historical data D:

$$\mathcal{L}(\Psi;D) = log\ p(D|\Psi). \tag{5}$$

We use Expectation-Maximization (EM) to find parameters Ψ that can maximize the log-likelihood of the historical data.

In the **E-step**, since there are two latent variables r and z in W^4, we update their joint expectation $p(r,z|d)$ according to Bayes rule as Equation 6.

$$p(r,z|d) = \frac{p(d,r,z)}{p(d)} = \frac{p(d,r,z)}{\sum_r \sum_z p(d,r,z)}. \tag{6}$$

In the **M-step**, we find the new Ψ that can maximize the log-likelihood as follows:

$$p(r|u,s) = \frac{\sum_{d\in D_{u,s}} \sum_z p(r,z|d)}{\sum_{d\in D_{u,s}} \sum_z \sum_{r'} p(r',z|d)}, \tag{7}$$

where $D_{u,s}$ is the collection of tweets written by user u on the day s. We will not explain $D_{(.)}$ unless necessary.

$$p(z|u,r) = \frac{\sum_{d\in D_u} p(r,z|d)}{\sum_{d\in D_u} \sum_{z'} p(r,z'|d)}, \tag{8}$$

$$\nu_{u,s,r} = \frac{\sum_{d\in D_{u,s}} \sum_z p(r,z|d)\cdot t_d}{\sum_{d\in D_{u,s}} \sum_z p(r,z|d)}, \tag{9}$$

$$\sigma^2_{u,s,r} = \frac{\sum_{d\in D_{u,s}} \sum_z p(r,z|d)\cdot td^2(t_d,\nu_{u,s,r})}{\sum_{d\in D_{u,s}} \sum_z p(r,z|d)}, \tag{10}$$

where $td(t_1,t_2)$ is the difference between time in a day t_1 and t_2, because the time in a day is cyclical. Note that for each region, we get two sets of ν and σ^2, and the two ν's are 12-hour apart from each other. For example, ν for 1:00 and 23:00 can be either 0:00 or 12:00, but the σ^2 value for 0:00 is much smaller than that for 12:00. Obviously, 0:00 is a better choice for the mean than 12:00. Thus, between the two sets, we choose the ν, σ^2 pair with the smaller σ^2 value.

Estimating $p(w|r)$ and $p(w|z)$ is not straightforward, because they are coupled by the sum in logarithm in the log-likelihood, *i.e.,* $log(\lambda\, p(w|z) + (1-\lambda)p(w|r))$. We solve this problem by applying Jensen's inequality [14]. Because logarithm is a concave function, we have:

$$\begin{aligned} log(\lambda\, p(w|z) + (1-\lambda)p(w|r)) \geq \lambda\, log(p(w|z)) + (1-\lambda) log(p(w|r)), \\ log(\kappa p(\ell|z) + (1-\kappa)p(\ell|r)) \geq \kappa\, log(p(\ell|z)) + (1-\kappa) log(p(\ell|r)). \end{aligned}$$

By substituting the above two Equations into Equation 5, we have a lower bound of the log-likelihood. By maximizing the lower bound, we have:

$$p(w|r) = \frac{\sum_{d\in D_w} \sum_z c(w,\boldsymbol{w}_d) p(r,z|d)}{\sum_{w'} \sum_{d\in D'_w} \sum_z c(w',\boldsymbol{w}_d) p(r,z|d)}, \tag{11}$$

$$p(w|z) = \frac{\sum_{d\in D_w} \sum_r c(w,\boldsymbol{w}_d) p(r,z|d)}{\sum_{w'} \sum_{d\in D'_w} \sum_r c(w',\boldsymbol{w}_d) p(r,z|d)}, \tag{12}$$

$$\boldsymbol{\mu}_{u,s,r} = \frac{\sum_{d\in D_{u,s}} \sum_z p(r,z|d)\cdot \boldsymbol{cd}_{\ell_d}}{\sum_{d\in D_{u,s}} \sum_z p(r,z|d)}, \tag{13}$$

$$\boldsymbol{\Sigma}_{u,s,r} = \frac{\sum\limits_{d\in D_{u,s}} \sum_z p(r,z|d)\cdot (\boldsymbol{cd}_{\ell_d} - \boldsymbol{\mu}_{u,s,r})^T (\boldsymbol{cd}_{\ell_d} - \boldsymbol{\mu}_{u,s,r})}{\sum\limits_{d\in D_{u,s}} \sum_z p(r,z|d)}, \tag{14}$$

$$p(\ell|z) = \frac{\sum_{d\in D_\ell} \sum_r p(r,z|d)}{\sum_{d\in D_\ell} \sum_{z'} \sum_r p(r,z'|d)}. \tag{15}$$

Please note that the lower bound technique has no impact on the inference of other parameters, since they are surrounded in different logarithms. As a result, the derivative with respect to a parameter does not involve with the others.

3.4 Complexity analysis

We now analyze the time complexity of the inference algorithm proposed in Section 3.3. For Equation 6 in the E-step, the time complexity is $O(K|D||R| + K|W||R|)$, where K, $|D|$, $|W|$, $|R|$ are the number of topics, the size of collection or number of tweets, the total number of words in the collection, and the number of regions for each user, respectively. The evaluation of Equation 6 requires to estimate $p(\ell|r)$. Theoretically, the time complexity to estimate $p(\ell|r)$ is $O(|U||R||L|)$. However, early pruning of locations faraway from r can be conducted. For example, when estimating $p(\ell|r)$ for a user in the United States, it is unnecessary to calculate $p(\ell|r)$ for locations in Singapore, since its value approximates to zero based on **Lemma 1**. A number of indexing techniques can be employed to achieve the early pruning of the search space $|L|$, such as R*-Tree [2], Quad-tree [10], etc. For M-step, the time complexity for Equations 7, 8, 9, 10, 13, 14, 15 is $O(K|D||R|)$, and is $O(K|W||R|)$ for Equations 11, 12. Thus, the time complexity for the M-Step is $O(T(K|D||R| + K|W||R| + |U||R||L|))$, where T is the number of EM iterations, which is set to 50 in our experiments.

4. APPLICATIONS

The proposed model W^4 has a variety of applications. We name some of them as examples:

Location prediction for tweet. Given a tweet with its text content, user id, and posting time, the task of *location prediction* is to predict the most likely location at which this tweet is posted. It has been shown [7,8] that geographical locations can be used to predict user's behavior, discover users' interest, and deliver location-based advertisement or content. However, it is reported that only 1%–2% of tweets have geographical locations explicitly attached. Hence, location prediction for tweets is an very important application.

A number of methods have been proposed for this task [9, 13, 15, 17, 24]. The studies [15, 17] build language models for each candidate location, and make prediction based on these language models. They are designed to predict location identifier for a text. Instead of predicting a location for a give text, the work [24] segments the world into grids, and employs supervised models, such as Naive Bayes, to predict grid for a given text. The recent proposal [13] presents a new approach for predicting geographic coordinates of a text from a user(See Section 2 for details).

Since W^4 incorporates both location identifiers and geographic coordinates, we can make both kinds of predictions for a text from a user, namely, predicting location identifiers [15, 17] and geographic coordinates [13]. Our method is also able to take the time factor into consideration.

Formally, given a user u, day s, time t, and words $\boldsymbol{w}_d$, a location ℓ (represented with both location identifier and geographic coordinates) is predicted by maximizing $p(\ell|u, s, t, \boldsymbol{w}_d)$. Specifically, we calculate $p(\ell|u, s, t, \boldsymbol{w}_d)$ for each candidate location ℓ as follows:

$$p(\ell|u, s, t, \boldsymbol{w}_d) = \frac{\sum_z \sum_r p(u, s, t, r, z, \boldsymbol{w}_d, \ell)}{\sum_z \sum_r \sum_{\ell'} p(u, s, t, r, z, \boldsymbol{w}_d, \ell')}, \quad (16)$$

where $p(u, s, t, r, z, \boldsymbol{w}_d, \ell)$ is computed as Equation 3.

Requirement-aware location recommendation. Location recommendation aims to recommend new locations for users. Previous studies only rely on users' historical visiting information [6, 25], neglecting the specific needs at a given time. W^4 is able to utilize both the time and the needs (in the form of short text), to make more accurate recommendation. Given a user u, day s, time t and words $\boldsymbol{w}_d$ that describe the need, the candidate locations are ranked by $p(\ell|u, s, t, \boldsymbol{w}_d)$, defined by Equation 16, and the top ranked ones are returned as results.

Activity prediction. W^4 is able to predict the activity of a user at a given time. Specifically, given a user u and time s and t, the words describing the activity are ranked by:

$$p(w|u, s, t) = \frac{\sum_z \sum_r p(u, s, t, r, z, w)}{\sum_z \sum_r \sum_{w'} p(u, s, t, r, z, w')}, \quad (17)$$

where $p(u, s, t, r, z, w) = p(u)p(s|u)p(r|u, s)p(t|u, s, r)p(z|u, r)p(w|r, z)$.

User prediction. User prediction aims to predict the likelihood of a user visiting a location at a given time. This could be very useful for merchants for planning purpose, or for them to target on specific costumers. Specifically, given location ℓ, day s, and time t, we rank candidate users by $p(u|\ell, s, t)$, which is calculated as follows:

$$p(u|\ell, s, t) = \frac{\sum_z \sum_r p(u, s, t, r, z, \ell)}{\sum_z \sum_r \sum_{u'} p(u', s, t, r, z, \ell)}, \quad (18)$$

where $p(u, s, t, r, z, \ell) = p(u)p(s|u)p(r|u, s)p(t|u, s, r)p(z|u, r)p(\ell|r, z)$. Note that previous studies on user mobility modeling (*e.g.,* [8]) can also be used for user prediction, if we use location and time as input, and find the user who can maximize the likelihood.

Location prediction for user. This task is to predict the place where a user stays at a given time. This would be useful for logistic planning, *e.g.,* to arrange a meeting with a user or a group of users, and location-based advertisement delivery. Formally, given a user u and time t, we aim to rank all candidate locations based on $p(\ell|u, s, t)$, which is calculated by:

$$p(\ell|u, s, t) = \frac{\sum_z \sum_r p(u, s, t, r, z, \ell)}{\sum_z \sum_r \sum_{\ell'} p(u, s, t, r, z, \ell')}. \quad (19)$$

Tweets recommendation. This task is to recommend tweets that are interested to a user based on the user's topic preferences, current location and time. Specifically, given user u, day s, time t, and location ℓ, we aim to rank tweets by considering $p(\boldsymbol{w}_d|u, s, t, \ell)$, where $\boldsymbol{w}_d$ is the word vector of a candidate tweets, and

$$p(\boldsymbol{w}_d|u, s, t, \ell) = \frac{\sum_z \sum_r p(u, s, t, r, z, \boldsymbol{w}_d, \ell)}{\sum_z \sum_r \sum_{\boldsymbol{w}'_d} p(u, s, t, r, z, \boldsymbol{w}'_d, \ell)}. \quad (20)$$

Table 1: Statistics of the two datasets

	WW	*USA*
Number of users	3,883	4,122
Number of locations	60,962	35,989
Number of tweets/messages	89,007	171,768

5. EXPERIMENTAL EVALUATION

We evaluate the proposed model in this section. Against several state-of-the-art baseline methods, we examine the accuracy of W^4 for the application of *location prediction for tweets* in Section 5.2. We present samples of the discovered topics and the mobility patterns of users in Section 5.3. Results of other example applications of W^4 are reported in Section 5.4.

5.1 Dataset

In our experiments, we use two real-world datasets, namely, *WW* dataset and *USA* dataset.

WW Dataset. Using the streaming API provided by Twitter[2], we collect a large volume of tweets with location information from November 1, 2012 to February 13, 2013. We refer to this dataset as *WW* (World-wide) dataset as the tweets are from users in different countries.

USA dataset. This dataset is the GeoText[3] (Geo-tagged Microblog Corpus) published by researchers from Carnegie Mellon University [9]. This dataset comprises messages from geo-located microblog users approximately in the United States. Each message is associated with its geographic coordinate. To map the geographic coordinates of each message to a location identifier, we crawl the geographic coordinates of locations in United States from Foursquare, and map the coordinates of each message to its nearest location.

For both datasets, we remove stop-words, and keep only the active users who visited at least 5 different locations. The statistics of the datasets after pre-processing is shown in Table 1. For each dataset, we randomly split the documents (tweets or messages) into three collections in proportion of 8:1:2 as the training set, development set, and testing set, respectively.

5.2 Location Prediction for Tweets

Given a tweet with its text content, user id, and posting time, the task of *location prediction* is to predict the most likely location at which this tweet is posted.

5.2.1 Evaluation Metrics

To evaluate the prediction performance of different models, we use two metrics, namely, prediction accuracy (Acc) and average error distance (Dis).

Prediction accuracy (Acc) is the percentage of tweets for which the predicted locations are exactly the true location among all tweets in the test set.

Average error distance (Dis) is the average of the Euclidian distance between the predicted geographic coordinates and the true geographic coordinates for all tweets in the test set.

[2] https://dev.twitter.com/docs/streaming-apis

[3] http://www.ark.cs.cmu.edu/GeoText/

Table 2: Comparison of baseline methods with W^3 and W^4

Factors in modeling	KL	TR	W^3	W^4
Who (User)	×	√	√	√
Where (Geo)	×	GlbR	PsnR	PsnR
When (Time)	×	×	×	√
What (Words)	√	√	√	√

Note that Acc and Dis are different—it is possible that the number of correctly predicted tweets is similar, but the wrongly predicted locations are deviated from the true locations very differently for different methods. Apparently, larger Acc and smaller Dis indicate better prediction performance.

5.2.2 Baseline methods

We compare with two baseline methods to evaluate the performance, which are the state-of-the-art models for predicting locations for text.

KL-divergence based method (KL) [15, 17]. This method builds language models (LM) for each candidate location during training. Given a test text, it computes the KL-divergence between the LM of the test text and the LM of each candidate location, and returns the most close location as the result.

Topic+Region (TR) [13]. This model captures the user preference over latent regions and topics. The location is generated from the Gaussian of regions, and words are generated based on the topic and region. This model represent locations as geographic coordinates. In addition, the latent regions in this model are not personal. Given a tweet from a user, TR predicts the geographic coordinates of the tweet, but cannot return the location identifier. Thus we cannot compute Acc for TR. In order to compare with other approaches in terms of Acc, we identify the location identifier for the predicted geographic coordinates by finding the nearest location to the coordinates.

Neither KL nor TR method makes use of the time factor in prediction. To study the performance of our model without time factor, we also use a simplified version of our proposed method as a baseline method.

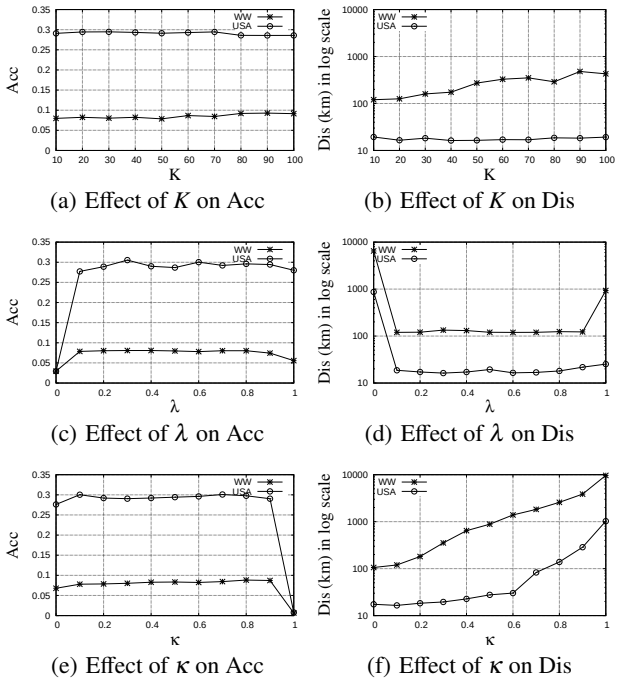
Who+Where+What (W^3). This W^3 method is based on similar inference modeling as our proposed model without considering the time factor (*i.e.,* time of a day and day of a week). Note that W^3 considers the similar set of aspects as does the TR model [13], but its modeling method is different from TR.

Who+Where+When+What (W^4). The differences between W^4 and other methods are summarized in Table 2, where "PsnR" and "GlbR" represent "using geographical information by estimating personal regions" and "using geographical information by estimating global regions for all users", respectively.

5.2.3 Parameter Setting and Tuning

We fix the number of personal regions as 2 for each user (*e.g.,* home region and work region), following the setting in [8]. Note that our model is capable of dealing with a larger number of personal regions. In our datasets, we notice the cases that a user may visit only one location at a region, or visit a region at only one time point. Such cases will result in problems like (i) errors in calculating the inverse of $\mathbf{\Sigma}_{u,s,r}$, or (ii) always getting zero value for $p(t|r)$. To avoid these problems, we set the minimum values for the determinant of $\mathbf{\Sigma}_{u,s,r}$ and $\sigma^2_{u,s,r}$ to be 1e − 16 and 1, respectively.

We set the three parameters in our model, namely, K: the number of topics, λ: the weight of $p(w|z)$, and κ: the weight of $p(\ell|z)$, by tuning them one by one on the development set. The default values for them are 60, 0.5, and 0.1, respectively. The tuning results on both datasets are reported in Figure 3. The impacts of varying the three parameters are discussed below.



(a) Effect of K on Acc (b) Effect of K on Dis

(c) Effect of λ on Acc (d) Effect of λ on Dis

(e) Effect of κ on Acc (f) Effect of κ on Dis

Figure 3: Tuning parameters for W^4

We first study the effect of number of topics K on the prediction performance. Figures 3(a) and (b) report the results of varying K from 10 to 100 on both datasets. Observe that K has almost no impact on Acc on both datasets. It has little impact on Dis on *USA* dataset. However, a larger K usually results in greater Dis on *WW* dataset. Recall that Acc and Dis are two different metrics, and Dis could be very different for different parameter settings even with similar Acc. We set K to 10 for *WW* data and 20 for *USA* data.

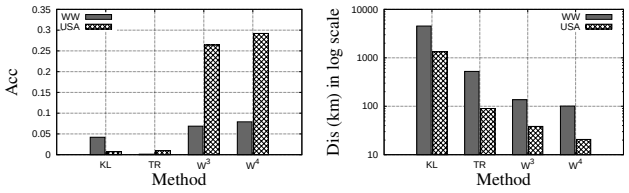
Next we tune λ. Observe from Figures 3(c) and (d), when $\lambda = 0$ or 1, Acc and Dis on both datasets are worse than other λ values between 0 and 1. This result shows that word variations of both regions and topics are important for prediction. We set λ to 0.6 for both datasets.

Finally we tune parameter κ. The results are shown in Figures 3(e) and (f). We observe that on both datasets, as κ is increased, Dis increases, but Acc keeps stable. However, Acc almost drops to zero at $\kappa = 1.0$, where the location selection is made only based on the topics ($p(\ell|z)$), and the region information ($p(\ell|r)$) of users is ignored. This is understandable since locations of all over the world can be returned as prediction results if the topic matches. At $\kappa = 0$, the prediction is purely based on the region of the user without taking into account the topics of the user—the predicted locations would be close to the mean location of the regions of the user. Finally, we set $\kappa = 0.1$ for both datasets.

5.2.4 Experimental results

We compare the prediction performance of the four methods (KL, TR, W^3, and W^4). The Dis and Acc of each method are reported in Figure 4. Note that only W^4 makes use of the time information in prediction.

As shown in Figure 4, W^4 outperforms the state-of-the-art baseline methods KL and TR significantly in terms of both Acc and Dis.



(a) Acc of methods (b) Dis of methods

Figure 4: Performance of all methods

W^4 outperforms KL in terms of Acc by 88.50% and 3953.04% on *WW* and *USA* datasets, respectively. In terms of Dis, compared with TR, W^4 reduces the average error distance by 80.73% and 77.02% on the two datasets, respectively.

KL is designed to predict the location label for short text. Because it does not exploit geographic coordinate information, its prediction performance in terms of Dis is much worse than other methods, *i.e.,* the average error distance of KL is much greater than those of the other methods. In addition, KL builds language model for locations based on the words posted by all users without considering the individuals' visiting history. In other words, it does not consider the preferences of individual users on locations. Moreover, the number of tweets posted at each location is small on average as observed from Table 1, and thus the language models of location are usually sparse, limiting the prediction performance of KL.

TR is designed to predict the geographic coordinates for short text. It returns the mean of the Gaussian distribution of the most likely latent region for a given tweet as the prediction result, but not the location identifier of the prediction. We observe that TR performs much better than KL in terms of Dis on both datasets. TR is based on topic models while KL adopts language models. Furthermore, TR incorporates the user preference information and the geographic coordinates information in its model. However, TR has the worst Acc among all methods, since the means of the global regions are less likely to be the exact locations of individuals' tweets.

Our model W^3 utilizes the same types of information as does TR, but it outperforms TR significantly. The reasons are two fold. First, the latent geographic regions in W^3 are personal while the latent geographic regions in TR is global for all the users. Hence, the regions in W^3 can describe individuals' mobility areas more precisely than the regions in TR. Second, both the location identifiers and the geographic information of locations are used by W^3 to enhance the prediction.

W^4 outperforms W^3 in terms of both measures. This is because W^4 incorporates the time factor in its model, which can further improves the prediction results. W^4 is capable of capturing the user's mobility patterns in terms of geographic, temporal, and activity aspects.

5.3 Sample Topics and Mobility Patterns

We take the model trained on *WW* dataset as an example to demonstrate the topics discovered by W^4.

We first randomly select 5 topics, and check their representative words. Specifically, for each topic z, we rank the words based on $p(w|z)$ and use the top-6 English words to represent each topic. The results are shown in Table 3.

For the ease of reading, we manually assign title for each topic. We find that, the representative words well reveal the semantic meaning of each topic.

Next, we randomly select a user, and look into the user's mobility patterns. We plot the two personal regions of the user in Figure 5, and the time patterns of each region in Figure 6. We assign the

Table 3: Representative words for topics

Topic	Representative words
Home	family fun offroad rental home love
Dinning	lunch dinner birthday breakfast drinks eat
Nightlife	night happy singing playing dance football
Work	working tonight coffee tired money friday
Holiday	christmas friends holiday merry celebrating choir

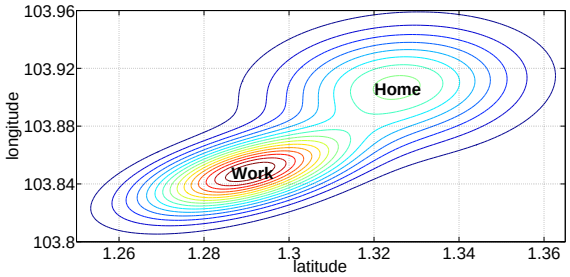


Figure 5: Personal regions

labels (*e.g.,* work and home) to the two regions based on the time of user visits. Figure 5 shows that the two geographic regions of the user are not far from each other. In addition, the contour lines of the work region are more close together than that of the home region, showing that the user usually stays in a small region at workplace, but visits a relatively larger range of places around her home.

From Figure 6, we observe that the user has different time patterns over the personal regions in weekdays and weekend. The time span that the user is more likely to stay in work region on weekends is much small than that on weekdays. In addition, the user is likely to spend more time in home region on weekends than that in weekdays.

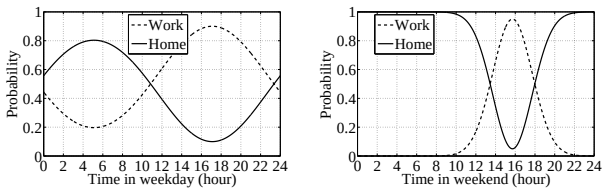
5.4 Results of Example Applications

In addition to location prediction for tweets, we implement another three applications, namely, activity prediction, user prediction and user's location prediction, and present their evaluation results in this subsection. We do not evaluate the location recommendation and tweet recommendation, because they require different datasets than what we use in our experiments.

Activity prediction Activity prediction returns the representative words describing user's activity at a given time. Using two different time as input (*i.e.,* 14:30 weekday, and 10:00 weekend), the top-6 (in terms of $p(w|u,s,t)$) English words returned by W^4 for a randomly selected user from *WW* data are shown in Table 4.

Observe that, in the weekday afternoon, the user's activity is more about work, taking a coffee break or resting. The user may also do body-building sometimes in the weekday afternoons. In the morning of weekends, the user stays at home for breakfast. Shopping and eating are also keywords for weekend mornings for the selected user.

User prediction. User prediction aims to predict the user who is most likely to visit a given location at a given time. We compare the performance of W^4 with a user mobility model PMM [8], on both datasets. Note that here we do not use the text of tweets, and thus PMM is applicable while the baseline approaches [13, 15, 17] for predicting locations of tweets take text as input and are not applicable here. For each tweet in test set, its time and location are used as input; if the predicted user is the true user of the tweet, it is a correct prediction. We employ prediction accuracy (Acc) as the evaluation metric, which shows the percentage of correct predictions.



(a) Weekday (b) Weekend

Figure 6: Region distribution over time

Table 4: Representative words for a user at a given time

time	words
14:30 weekday	break work coffee resting gym international
10:00 weekend	good morning home breakfast shopping eat

The results are reported in Table 5. In user prediction, W^4 outperforms PMM by 21.62% and 45.81% on the two datasets, respectively. Potential reasons are two-fold: on the one hand, we use a new way to calculate the probability of latent regions at a given time, which is different from the way used in PMM; on the other hand, our model exploits both the functional and graphical coordinate information of locations, while PMM only utilizes the latter.

Location prediction for user. This task aims to predict the location at which a given user is most likely stay at a given time. For each tweet in test set, its time and user are used as input; if the predicted location is the true location of the tweet, it is a correct prediction. We still evaluate the performance using prediction accuracy. The experimental results are reported in Table 6. For this task, the PMM method is also used as the baseline, where the input has no text. The results show that our method outperforms the baseline method significantly for similar reasons discussed earlier.

6. CONCLUSION AND FUTURE WORK

The large availability of geo-tagged tweets enables us to study individuals' mobility behaviors from four factors, namely user, geographic information, time, and activity. Unfortunately, none of the previous studies considers all of them. In this paper, we propose a probabilistic generative model W^4, which is capable of capturing the four factors jointly, and providing a comprehensive description of user mobility behavior. We evaluate the performance of W^4 for several applications on two real-world datasets, and the experimental results show that the proposed method outperforms state-of-the-art baselines significantly for these applications.

In the future, we aim to exploit the proposed model for other potential applications. In addition, it will be interesting to incorporate social information into the model.

7. ACKNOWLEDGEMENTS

This work is supported in part by a grant awarded by a Singapore MOE AcRF Tier 2 Grant (ARC30/12), a Singapore MOE AcRF Tier 1 Grant (RG66/12), and a grant awarded by Microsoft Research Asia. Quan Yuan would like to acknowledge the Ph.D. grant from the Institute for Media Innovation, Nanyang Technological University, Singapore.

Table 5: User prediction Acc of PMM and W^4

Acc	*WW*	*USA*
PMM	0.4163	0.4021
W^4	**0.5063**	**0.5863**

Table 6: Location prediction Acc of PMM and W^4

Acc	*WW*	*USA*
PMM	0.0423	0.1102
W^4	**0.0776**	**0.2953**

8. REFERENCES

[1] S. Bauer, A. Noulas, D. O. Seaghdha, S. Clark, and C. Mascolo. Talking places: Modelling and analysing linguistic content in foursquare. In *SocialCom/PASSAT*, pages 348–357, 2012.

[2] N. Beckmann, H.-P. Kriegel, R. Schneider, and B. Seeger. The r*-tree: An efficient and robust access method for points and rectangles. In *SIGMOD Conference*, pages 322–331, 1990.

[3] C. M. Bishop and N. M. Nasrabadi. *Pattern Recognition and Machine Learning*. *J. Electronic Imaging*, 16(4):049901, 2007.

[4] D. Brockmann, L. Hufnagel, and T. Geisel. The scaling laws of human travel. *Nature*, 439(7075):462–5, 2006.

[5] P. W. Chaoming Song, Tal Koren and A.-L. Barabási. Modelling the scaling properties of human mobility. *Nature Physics*, Sep 2010.

[6] C. Cheng, H. Yang, I. King, and M. R. Lyu. Fused matrix factorization with geographical and social influence in location-based social networks. In *AAAI*, 2012.

[7] Z. Cheng, J. Caverlee, K. Lee, and D. Z. Sui. Exploring millions of footprints in location sharing services. In *ICWSM*, 2011.

[8] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *KDD*, pages 1082–1090, 2011.

[9] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing. A latent variable model for geographic lexical variation. In *EMNLP*, pages 1277–1287, 2010.

[10] R. A. Finkel and J. L. Bentley. Quad trees: A data structure for retrieval on composite keys. *Acta Inf.*, 4:1–9, 1974.

[11] M. C. Gonźĺez, C. A. Hidalgo, and A. L. Barabasi. Understanding individual human mobility patterns. *Nature 453, 479-482 (2008)*, Jun 2008.

[12] Q. Hao, R. Cai, C. Wang, R. Xiao, J.-M. Yang, Y. Pang, and L. Zhang. Equip tourists with knowledge mined from travelogues. In *WWW*, pages 401–410, 2010.

[13] L. Hong, A. Ahmed, S. Gurumurthy, A. J. Smola, and K. Tsioutsiouliklis. Discovering geographical topics in the twitter stream. In *WWW*, pages 769–778, 2012.

[14] J. Jensen. Sur les fonctions convexes et les inĺegalitĺes entre les valeurs moyennes. *Acta Mathematica*, 30:175–193, 1906.

[15] S. Kinsella, V. Murdock, and N. O'Hare. "i'm eating a sandwich in glasgow": modeling locations with tweets. In *SMUC*, pages 61–68, 2011.

[16] C. Li, A. Sun, and A. Datta. Twevent: segment-based event detection from tweets. In *CIKM*, pages 155–164, 2012.

[17] W. Li, P. Serdyukov, A. P. de Vries, C. Eickhoff, and M. Larson. The where in the tweet. In *CIKM*, pages 2473–2476, 2011.

[18] Q. Mei, C. Liu, H. Su, and C. Zhai. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *WWW*, pages 533–542, 2006.

[19] A. Ritter, Mausam, O. Etzioni, and S. Clark. Open domain event extraction from twitter. In *KDD*, pages 1104–1112, 2012.

[20] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *WWW*, pages 851–860, 2010.

[21] S. Sizov. Geofolk: latent spatial semantics in web 2.0 social media. In *WSDM*, pages 281–290, 2010.

[22] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[23] C. Wang, J. Wang, X. Xie, and W.-Y. Ma. Mining geographic knowledge using location aware topic model. In *GIR*, pages 65–70, 2007.

[24] B. Wing and J. Baldridge. Simple supervised document geolocation with geodesic grids. In *ACL*, pages 955–964, 2011.

[25] M. Ye, P. Yin, W.-C. Lee, and D. L. Lee. Exploiting geographical influence for collaborative point-of-interest recommendation. In *SIGIR*, pages 325–334, 2011.

[26] Z. Yin, L. Cao, J. Han, C. Zhai, and T. S. Huang. Geographical topic discovery and comparison. In *WWW*, pages 247–256, 2011.