Who, Where, When and What: a Non-parametric Bayesian Approach to Context-aware Recommendation and Search for Twitter Users

Quan Yuan, Gao Cong, Kaiqi Zhao, Zongyang Ma, and Aixin Sun, Nanyang Technological University

Micro-blogging services and location-based social networks, such as Twitter, Weibo, and Foursquare, enable users to post short messages with timestamps and geographical annotations. The rich spatial-temporal-semantic information of individuals embedded in these geo-annotated short messages provides exciting opportunity to develop many context-aware applications in ubiquitous computing environments. Example applications include contextual recommendation and contextual search. To obtain accurate recommendations and most relevant search results, it is important to capture users' contextual information (*e.g.,* time and location) and to understand users' topical interests and intentions. While time and location can be readily captured by smartphones, understanding user's interests and intentions calls for effective methods in modeling user mobility behavior. Here, user mobility refers to *who visits which place at what time for what activity*. That is, user mobility behavior modeling must consider user (Who), spatial (Where), temporal (When), and activity (What) aspects. Unfortunately, no previous studies on user mobility behavior modeling have considered all of the four aspects jointly, which have complex interdependencies. In our preliminary study, we propose the first solution named W^4 (short for **W**ho, **W**here, **W**hen, and **W**hat) to discover user mobility behavior from the four aspects. In this article, we further enhance W^4 and propose a non-parametric Bayesian model named EW^4 (short for **E**nhanced W^4). EW^4 requires no parameter tuning and achieves better results over W^4 in our experiments. Given some of the four aspects of a user *e.g.,* time, our model is able to infer information of the other aspects *e.g.,* location and topical words. Thus, our model has a variety of context-aware applications, particularly in contextual search and recommendation. Experimental results on two real-world data sets show that the proposed model is effective in discovering users' spatial-temporal topics. The model also significantly outperforms state-of-the-art baselines for various tasks including location prediction for tweets and requirement-aware location recommendation.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms: Algorithms, Management, Experimentation

Additional Key Words and Phrases: Recommendation and Search, Context-aware, Geographical Topic Modeling, Requirement-aware Location Recommendation, Spatio-Temporal, Graphical Model, Twitter

1. INTRODUCTION

With the prevalence of 3G/4G technology, people are able to access Internet anytime anywhere via smartphones, tablets, wearable devices, *etc.* Mobile Internet service has become an indispensable part of people's daily life. For example, people are willing to share opinions, moods, and activities as tweets (short messages with maximum length of 140 characters) on Twitter, and share their experience and tips as check-ins on Foursquare, a location-based social network. As a result, a sheer amount of user-generated content (UGC) has been accumulated. As of December 2013, the

Portions of this work appeared in Yuan et al. [2013b].
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.
Author's addresses: School of Computer Engineering, Nanyang Technological University, Singapore 639798

DOI : http://dx.doi.org/10.1145/0000000.0000000

number of tweets was increased at the speed of 500 million per day. Especially, most UGC is in the form of short text, such as tweets, Facebook status, Foursquare shouts (short messages associated with check-ins) and tips.

As more and more devices for Internet access are GPS enabled, a large portion of user generated short text has been associated with spatial information. For example, Twitter users can specify the locations for their tweets to indicate the surroundings of posting, where the locations may be in the form of the latitude and longitude coordinates, or venues with semantic meanings, *e.g.,* Times Square. While the coordinates can be captured by GPS devices, the exact venues can be specified explicitly by users, detected by mobile devices, or annotated by geo-tagging tools. It has been reported that about 2.7% tweets contain geographic information about users' current surroundings according to a report dated on June 2013[1]. Besides geo-annotated tweets, the check-ins made by Foursquare users also contain the physical locations visited by the users. In addition to the spatial information embedded in the locations and the semantic information carried by the short text, the user-generated messages also contain temporal information, because the timestamp of each message is readily captured by the message posting services.

The availability of short messages with rich spatial-temporal-semantic information of individuals in large amount makes it possible to design various context-aware applications in ubiquitous computing environments, such as contextual recommender systems and contextual search. Recommender systems are designed to generate a list of recommended items for users, and search engines take keywords as input and present the most relevant results for users. For both applications, in addition to time and location, understanding users' interests and intentions are the pivots to achieve good accuracy.

Over the past two decades, significant research effort has been devoted to developing algorithms to improve the recommendation and search performance, and a number of effective methods have been proposed, such as matrix factorization for recommendation [Koren et al. 2009], and nature language processing techniques for search. In addition to this, more researchers have turned to make use of additional information to mine user interests for more accurate personalized results. Generally there are two approaches:

— Exploiting contextual information. Contextual information, such as location and time, reveals the surrounding environment and the current situations of users, which might be correlated with users' interests and intentions. For example, consider a white-collar who submits a recommendation request on weekday afternoon at her office. Then the recommender system can infer her interest based on the contextual information, namely, the time "weekday afternoon" and the location "office", and recommend her a coffee house near the office rather than a pub. Note that the contextual information can either be captured by mobile devices implicitly, or be specified by users explicitly.
— Mining user mobility behavior from historical UGC. User mobility behavior consists of four aspects, namely, user aspect (***who*** is the user), spatial aspect (***where*** does the user go?), temporal aspect (***when*** does the user visit a place?), and activity aspect (***what*** does the user do?). An example behavior pattern is: a white-collar mostly stays at her office on weekday afternoons, and likes to visit a coffee house near her office. Understanding user mobility behavior enables applications to capture user interests over time and locations, which is helpful in refining the list of recommended items and the list of search results.

We believe that the contextual information and user mobility behavior compliment each other, and the two together help us target users' interests and intentions. Without understanding user mobility behavior, the results would not be personalized; without contextual information, the results would not be specific to current environment and situations. In fact, user mobility behavior reflects the long-term interests of users, while contextual information helps discover users' short-term interests at the query time. Recall the white-collar who likes to go to coffee houses on weekday afternoons. Suppose she always visits pubs at CBD (central business district) on weekend evenings, then her

[1] http://irevolution.net/2013/06/09/mapping-global-twitter-heartbeat/

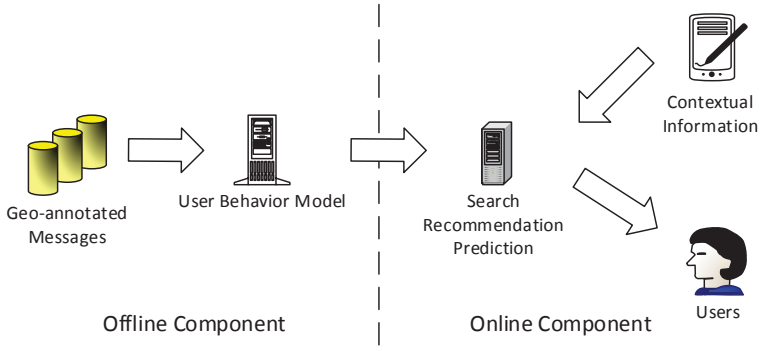


Fig. 1. High-level overview of the system architecture

preference to coffee houses and pubs at different time and locations is her long-term interest, which can be discovered by user mobility behavior models. Suppose she submits a recommendation request on a weekend evening at CBD, then the spatial and temporal contextual information enables us to predict her short-term interest in pubs.

Unfortunately, existing proposals on recommender systems and search engines have not fully utilized the contextual information and the user mobility patterns, and most of them neglect the valuable contextual information, *i.e.,* recommender systems recommend items solely based on the user-item purchase or rating matrix. Although several models have been proposed to model user behavior and are able to make context-aware recommendations or predictions, they cannot jointly model the four aspects of user behavior, *e.g.,* who, where, when, and what. For example, the studies [Hong et al. 2012; Ahmed et al. 2013] focus on the geographic location and activity aspects of users, but do not model the temporal aspect. As a result, they cannot take time as an additional piece of contextual information to achieve better recommendation accuracy.

In this article, we propose a framework that models user mobility behavior for context-aware applications. As shown in Figure 1, our proposed framework consists of the *offline* and *online* components. The offline component takes geo-annotate messages as input, and trains a user behavior model that incorporates the interactions of all the four aspects in an integrated manner. The online component takes in the contextual information of users and returns different recommendation and search results to users for different applications. The main technical challenge of this article lies in the offline component for the user mobility behavior modeling. It is however difficult to develop such a model, because the interdependencies among the four aspects and the role played by each are unclear. What's more, the parameter estimation for the model would be very complex. To model user mobility behavior from the spatial, temporal, and activity aspects, we take the following intuitions into account:

(1) An individual's mobility usually centers at several personal geographical regions, *e.g.,* home region and work region [Cho et al. 2011] and users tend to visit the places within these regions. In addition, the number of personal regions is user-specific, *e.g.,* some users may have additional regions for shopping and weekend activities.
(2) The probability that a user stays at a given region is affected by the day of the week, *e.g.,* users are more likely to stay at the work regions on weekdays than weekends. Moreover, given a region, users may have different temporal patterns on different days (weekdays or weekends), *e.g.,* a user may visit her shopping region in weekend afternoons, but in weekday evenings.
(3) Users engage in different activities at different places, and the topics of a user at a place are influenced by both the user's personal topic preference and the region where the user stays. For example, a student who is interested in topics like reading and shopping will concentrate on the shopping topic rather than reading topic when she is at Times Square.

(4) When choosing a location to visit, a user will consider both her personal topic preference and the geographic coordinates, *i.e.,* whether the location matches her topic preference, and whether the location is within her current region of stay.
(5) Different regions and topics lead to different word variations. Thus, the words used in a tweet posted by a user at a location are influenced by the user's current region and her topic preference, which in turn reflect the user's activity. For example, if a user is shopping at her home region, the words she would use are more likely to be related to both the shopping topic and the home region, such as "grocery", "family", *etc.*

Based on these intuitions, in our preliminary work[Yuan et al. 2013b], we propose a probabilistic Latent Semantic Analysis (pSLA) [Hofmann 1999] based model W^4 (short for **W**ho, **W**here, **W**hen, and **W**hat) to characterize the users' mobility behaviors from the spatial, temporal and activity aspects in a principled framework. W^4 can discover spatial-temporal topics, and identify personal geographical regions and time-aware user interests from the geo-annotated messages. However, W^4 is built under the strong assumptions that every user has the same number (*i.e.,* 2) of personal regions, and makes the same trade-offs between topics and personal regions when generating locations and words. In this article, we relax these assumptions, and develop an enhanced model named EW^4 (short for **E**nhanced **W**ho, **W**here, **W**hen, and **W**hat). In EW^4, all parameters (*e.g.,* the number of personal regions of each user, and the trade-offs between topics and regions) can be automatically learned from the training data. In addition, as a Hierarchical Dirichlet Process (HDP) [Teh et al. 2006] based model, EW^4 is less sensitive to the overfitting problem than W^4.

Given information for some of the four aspects (*i.e.,* user, spatial, temporal and activity aspects), both W^4 and EW^4 are able to infer the other aspects, where the activity aspect is represented by words. Thus, the proposed models can be served as the offline component of the framework for various context-aware applications. The following are example applications among many others.

— *Location-aware search*. It has been reported that the location of a user may reveal their search interests and intentions, and can be exploited to improve the relevance of search results [Jones et al. 2008; Bennett et al. 2011; Yan et al. 2014]. Our models are able to detect users' topical preference at different locations (*e.g.,* home, workplace, *etc.*), and refine the search results based on the users' current locations.
— *Location-prediction*. Although GPS-enabled devices can capture users' coordinates, the location services may be switched off by users for energy saving or other purposes. Moreover, GPS devices usually do not work well in indoor environments. Because our models jointly model location, time and semantic information, locations of users can be inferred based on their query keywords and time.
— *Time and requirement aware location recommendation*. It has been reported that time is an important factor that influences user mobility, and incorporating temporal influence can significantly improve recommendation accuracy [Yuan et al. 2013a, 2014]. In addition, when querying for recommendations, users may have specific needs expressed in short text, *e.g.,* "chessy pizza", "budget shopping mall". Exploiting users' specific needs can definitely improve the accuracy of the recommendation results. Our models are able to make use of time factor and specific needs as additional evidences to better understand users' intentions, and make more accurate recommendations.

Besides these applications, W^4 and EW^4 can also be applied in user profiling, location identification, topic tracking, *etc.* In the experiments, we evaluate their effectiveness in various applications including location prediction for tweets (with or without time), requirement-aware location recommendation for individual users, location prediction for a target user at a given time, and predicting the users who will visit a given location at a given time. Experimental results show that W^4 outperforms existing approaches [Li et al. 2011; Cho et al. 2011; Hong et al. 2012; Hu and Ester 2013; Ahmed et al. 2013] for these applications, and the enhanced version EW^4 further outperforms W^4 significantly.

The contributions of this work are summarized as follows:

— We develop a new probabilistic generative model EW^4 to model users' mobility behavior from user, spatial, temporal and activity aspects in an integrated way. The model enables us to discover spatial-temporal topics for individual users, and to make context-aware recommendations and search.
— We propose a new inference algorithm for estimating the model parameters.
— We define a new problem, namely, requirement-aware location recommendation, which aims at recommending locations for a target user based on her specific requirement (and time).
— Experimental results on two real-world data sets show that EW^4 is capable of identifying interesting spatial-temporal topics for users. The results also show that EW^4 significantly outperforms the state-of-the-art baselines for various applications.

The rest of this paper is organized as follows. We survey the related studies in Section 2, and characterize users' mobility behavior in Section 3. Section 4 introduces the proposed model for the offline component, namely, EW^4. The algorithm for parameter estimation is presented in Section 5. Potential applications for the online component are discussed in Section 6. The experimental results are presented in Section 7. Section 8 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 [Wang et al. 2007; Hao et al. 2010] 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.* [Wang et al. 2007] 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.* [Hao et al. 2010] mine the location-representative topics from travelogues using an LDA-based model. In the model, a travelogue is split into several segments along with locations, and the words in each segments are generated either from local topics or global topics. Comparing to travelogues, tweets are very short, and each tweet can be associated with only one location.

Second, other proposals [Eisenstein et al. 2010; Sizov 2010; Yin et al. 2011] 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 practice. Eisentein *et al.* [Eisenstein et al. 2010] 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 [Sizov 2010] 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 grows. In its subsequent work [Kling et al. 2014], Kling *et al.* propose a multi-Dirichlet process (MDP) based model to detect non-Gaussian geographical topics. Yin *et al.* [Yin et al. 2011] propose a probabilistic Latent Semantic Analysis (pLSA) based model to discover geo-

graphical 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.* [Mei et al. 2006] 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 [Bauer et al. 2012], where a city is divided into grids. Compared with the models [Mei et al. 2006; Bauer et al. 2012], our model considers more aspects:1) the models [Mei et al. 2006; Bauer et al. 2012] do not consider the user information at all; 2) it either does not consider the geographic property of locations [Mei et al. 2006], or does not consider the functions of locations [Bauer et al. 2012]; 3) they only consider discretized time.

There are also several studies on extracting events from Twitter stream [Li et al. 2012; Ritter et al. 2012], which exploit the temporal (When) and activity (What) information, and some work even considers the geographic aspect (Where) [Sakaki et al. 2010]. 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.* [Brockmann et al. 2006] find that human mobility behavior can be approximated by the continuous-time random-walk model. Gonzlez *et al.* [Gonzalez et al. 2008] 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.* [Song et al. 2010b,a] 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.* [Cho et al. 2011] 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: There are several studies on modeling the geographic (Where) and activity (What) aspects for individuals (Who).

Hong *et al.* [Hong et al. 2012] 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 depending on both the region and the user. The topic, together with the region, generates the words of a tweet; the region alone generates the coordinates based on its Gaussian distribution over coordinates.

In their subsequent work [Ahmed et al. 2013], the authors consider the relations between regions, and propose a Chinese Restaurant Franchise (nCRF) based model to study users' geographical topics. Specifically, regions in the model are organized hierarchically, where regions in upper levels geographically encompass those in lower levels, and are more diverse in terms of topics. A tweet defines a path from the root region to the leaf region, where its coordinates are sampled based on the Gaussian distribution of the leaf region, and its text content is generated based on the topics and language model of the leaf region.

Different from our work, these two studies does not consider the temporal aspect. In addition, the regions in the two models 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 methods in the two studies fail to consider the semantic information of individual locations in the same region.

Hu *et al.* [Hu and Ester 2013] propose a model that considers both the coordinates and semantic information of locations. In this model, the topic and region of a tweet is drawn based on both global and user-specific topic and region distributions, respectively. The topic further generates the text content, where the region and topic together determine the location of the tweet. However, this work does not consider the temporal aspect, and does not model the interaction between regions and topics.

In summary, none of existing studies aim to model the four aspects (Who, Where, When, and What). In contrast, our two models jointly consider the four aspects. In our proposed models, different users have different personal regions, and the regions at which a user stay is influenced by both the day of a week and the time of a day. Each personal region has a Gaussian distribution over coordinates, a topic distribution and a word distribution. Each tweet has a topic, which is drawn based on the topic distribution of the personal region, and the topic, together with the region, generate the location and words of the tweet. Note that this is the first work that models the coordinates and functions of locations simultaneously, and it can capture both the geographic region and functional information of locations.

Who Where When What: The W^4 model proposed in our paper [Yuan et al. 2013b] is the first work that jointly models who, where, when and what aspects. It models the generative process of a geo-tagged tweet as follows: a region r is first drawn based on the author u's distribution over personal regions at time t, and then a topic z is drawn based on u's topic distribution at r. Finally, the location of the tweet is drawn from a weighted combination of topic z's distribution over location identifiers and region r's Gaussian distribution over coordinates, and the words of the tweet are drawn from a weighted combination of the topic z's and region r's word distributions.

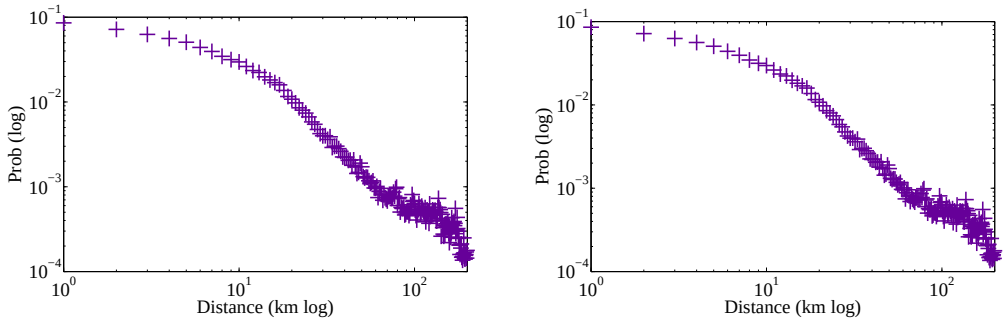
This article substantially extends the W^4 model in the following aspects. First, the W^4 model is based on probabilistic Latent Semantic Analysis (pLSA) model, and is sensitive to overfitting problem. In addition, several parameters of W^4 need manual tuning. In contrast, the EW^4 model is based on HDP model and is a non-parametric Bayesian model, in which parameters can be automatically learned from the training data. Second, in the W^4 model, all users have the same number of personal regions. In contrast, in EW^4, users can have different numbers of personal regions, which is modeled by the Chinese Restaurant Process (CRP). Third, the W^4 model assumes every user makes the same trade-off between regions and topics when selecting locations and words, while EW^4 learns the weight of each part from training data. Fourth, we include a new task in the experiments, namely, time and requirement-aware location recommendation, which aims at recommending a list of locations for the target users to visit at the target time based on their specific requirements. Finally, as to be shown by our experimental studies, the EW^4 model performs much better than the W^4 model for various applications.

3. CHARACTERISTICS OF INDIVIDUAL MOBILITY

In this section, we study the characteristics of individual user mobility pattern on two data sets, namely, World-Wide tweets collection (WW) and microblogs collection from USA (USA). More details about the two data sets are reported in Section 7.

We first examine the effect of spatial distance to users' mobility. Specifically, for each user, we calculate the distance between every pair of her visited locations. Then, we aggregate the results of all users and plot the number of check-ins as a function of distance in Figure 2. From the figures, we observe that the probability distribution follows a power law function on both data sets. This observation is consistent with the observation made by [Ye et al. 2011]. The results show that users are more likely to visit locations close to their visited locations, and thus the locations visited by a user form several spatial clusters.

To better illustrate the spatial clusters, we randomly select a user from the data set and plot her visited locations in Figure 3. The locations visited on weekdays and weekends are plotted in different colors. From the figure, we observe that the locations visited by the user can be grouped into



(a) WW (b) USA

Fig. 2. Distribution of distance between pairs of check-ins

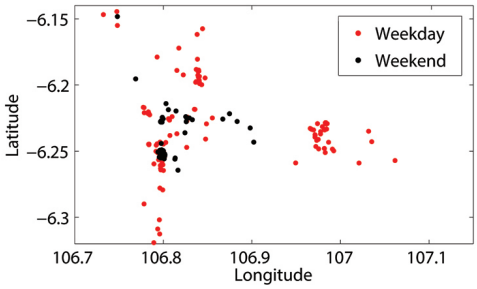
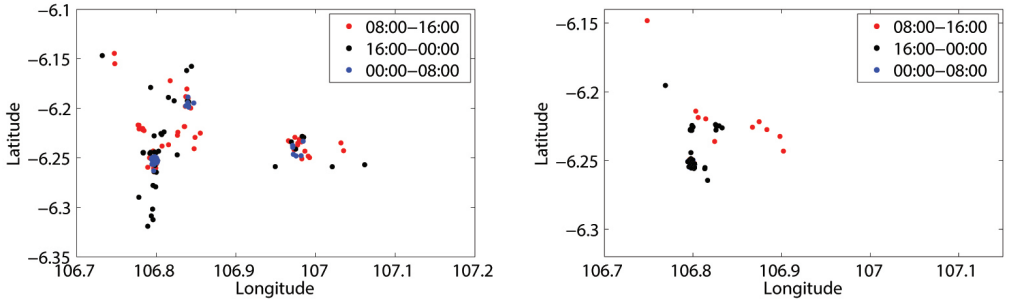


Fig. 3. Visited locations of a user plotted on a map



(a) Weekdays (b) Weekends

Fig. 4. Visiting time of locations in weekdays and in weekends

several geographical regions, and the user visits a region with different probabilities on weekdays and weekends.

Next, we analyze the temporal pattern of users' mobility. Specifically, we divide the check-ins of a randomly selected user into two sets based on the check-in days (*i.e.,* weekdays and weekends). For each set, we plot the visited locations on the map and distinguish three visiting time slots by using three different colors in Figure 4. From the figures, we make an observation that the visiting time of locations in a region is different on weekdays and weekends.

4. PROPOSED MODEL

In this section, we introduce the offline component of our framework that models user mobility behavior. Specifically, we first describe the modeling intuitions in Section 4.1, and then introduce the EW^4 model in Section 4.2.

4.1. Intuitions and Notations

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

Intuition 1. An individual's mobility usually centers at several personal geographical regions, *e.g.,* home region and work region [Cho et al. 2011]. The number of personal regions is user-dependent, *e.g.,* users who are interested in outdoor activities may have hiking regions, and users who are interested in shopping may have shopping regions.

Intuition 2. A user may visit a region with different probabilities on weekdays and weekends (Section 3), *e.g.,* she may go to work region on weekdays rather than on weekends. In addition, the visiting time of a region is influenced by the day (Section 3), *e.g.,* a user is more likely to stay at home region in the evenings of weekdays, as well as daytime on weekends.

Intuition 3. The topics of a user at a place are influenced by both the user's personal topic preference 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, then 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.

Intuition 4. When choosing a location to visit, a user will consider both the topic requirement and the region where the user stays. Intuitively, a user tends to visit nearby locations within her current region of stay that meet her requirement (*e.g.,* for meal). In addition, different users may make different trade-offs between the topic and region factors, *e.g.,* comparing to those without cars, the users who have cars may treat the region with less importance, because it is much easier for them to drive to the locations they want to visit.

Intuition 5. Different regions and different topics lead to different language variations, which in turn reflect the users' activities. Therefore, the words in a user's tweets 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 more likely to be related to both the shopping topic and the home region, such as "grocery", "family", *etc.* In addition, the weight of each part is also user-specific.

We consider each user u has several personal regions, denoted by $\{r_{u,0}, r_{u,1}, ..., r_{u,|R_u|}\}$, where $|R_u|$ is the number of regions of user u. The personal regions are estimated based on the locations of all geo-tagged tweets from user u. We model a 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. A region r is modeled by a bi-variant Gaussian distribution over the latitude and longitude, parameterized by the mean vector $\boldsymbol{\mu}_r$ and covariance matrix $\boldsymbol{\Lambda}_r^{-1}$. Note that we use r to represent a region (*i.e.,* any one of the personal regions) when the semantic is clear.

We model time t in a day as a continuous variable in $\langle hh : mm : ss \rangle$ format, and categorize days into two classes, namely, weekdays and weekends. Specifically, we use $s \in S = \{0, 1\}$ 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 cyclic 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.

We consider a tweet d is a five-tuple $d_i = \{u_i, \ell_i, \boldsymbol{w}_i, t_i, s_i\}$, where u_i denotes the user or the author of the tweet; ℓ_i, t_i, and s_i denote the location, the time in a day, and the day of the week, as described earlier; $\boldsymbol{w}_i$ are the words in tweet d_i. For easy presentation, we use D, U, and L to denote the collections of tweets, users, and locations respectively. The word vocabulary is denoted by V.

Table I. Symbols

Symbol	Description
U, L, S, W, D	user set, location set, day set {weekday, weekend}, vocabulary set, tweets set
Z, R, R_u	topic set, region set of all users, region set of user u, where $R = \bigcup_{u \in U} R_u$
$\|(\cdot)\|$	the number of elements in set $(\cdot)$
u, ℓ, w, s, t	user $u \in U$, location $\ell \in L$, word $w \in W$, day of a week $s \in S$, time of a day $\langle hh : mm : ss \rangle$
d_i, z_i, r_i, $\boldsymbol{w}_i$	the i^{th} tweet in D, the topic, region and the words in tweet d_i
$r_{u,j}$	the j^{th} region of user u, where $1 \leq j \leq \|R_u\|$
c_i^L, c_i^W	the location and word switches for tweet i
$c_{w,\boldsymbol{w}}$	number of times the word w appears in $\boldsymbol{w}$
$\{t\}_r$	the collection of time of the tweets that are assigned to region r
$td(t_1, t_2)$	the difference between time t_1 and t_2 in a day
G_0	global probability measure over topic space with mixing proportion $\boldsymbol{\tau}$
G_r	region-specific measure over topic space with mixing proportion $\boldsymbol{\theta}_r$
$\boldsymbol{\tau}$	global multinomial distribution of topics
$\boldsymbol{\theta}_r$	multinomial distribution of topics specific to region r
$\boldsymbol{\psi}_{u,s}$	multinomial distribution of regions specific to user u on day s
$\boldsymbol{\phi}_z^{ZL}$	multinomial distribution of locations specific to topic z
$\boldsymbol{\phi}_z^{ZW}$	multinomial distribution of words specific to topic z
$\boldsymbol{\phi}_r^{RW}$	multinomial distribution of words specific to region r
$\boldsymbol{\xi}_u^L$	Bernoulli distribution specific to user u for sampling the binary switch c^L
$\boldsymbol{\xi}_u^W$	Bernoulli distribution specific to user u for sampling the binary switch c^W
$\boldsymbol{\mu}_r$, $\boldsymbol{\Lambda}_r$	mean, precision matrix of Gaussian distribution over geographic coordinates specific to region r
$\nu_{r,s}$, $\lambda_{r,s}$	mean and precision of Gaussian distribution over time specific to region r in day s
γ	parameter of the prior of $\boldsymbol{\tau}$
α	concentration parameter for G_r
β	parameter for Chinese Restaurant Process for $\boldsymbol{\psi}$
η, χ, ζ	Dirichlet prior vector for $\boldsymbol{\phi}^{ZL}$, $\boldsymbol{\phi}^{ZW}$, and $\boldsymbol{\phi}^{RW}$
$\boldsymbol{o}$, $\boldsymbol{\delta}$	Beta priors for $\boldsymbol{\xi}^L$ and $\boldsymbol{\xi}^W$, where $\boldsymbol{o} = \{o_0, o_1\}$ and $\boldsymbol{\delta} = \{\delta_0, \delta_1\}$
$\boldsymbol{\mu_0}$, κ_0, υ_0, $\boldsymbol{\epsilon_0}$	Normal-Wishart prior for $\boldsymbol{\mu}$, and $\boldsymbol{\Lambda}$
ν_0, ι_0, ρ_0, λ_0	Normal-Gamma prior for ν, and λ
$n_{s,r,\neg i}^{SR}$	number of times region r is assigned to day s, excluding tweet i
$n_{r,z,\neg i}^{RZ}$	number of times topic z is assigned to region r, excluding tweet i
$n_{z,\ell,\neg i}^{ZL}$	number of times location ℓ is assigned to topic z, excluding tweet i
$n_{z,w,\neg i}^{ZW}$	number of times word w is assigned to topic z, excluding tweet i
$n_{r,w,\neg i}^{RW}$	number of times word w is assigned to region r, excluding tweet i
$n_{u,(\cdot),\neg i}^{UCL}$	number of times switch $c^L = (\cdot)$ is assigned to user u, excluding tweet i
$n_{u,(\cdot),\neg i}^{UCW}$	number of times switch $c^W = (\cdot)$ is assigned to user u, excluding tweet i
m_z	number of tweets that are assigned to topic z

That is, $d \in D$, $u \in U$, $\ell \in L$, and each word in $\boldsymbol{w}_i$ belongs to V. All notions used in this article are shown in Table I.

4.2. Generative Process

EW^4 generates the day, time, words, and location for each tweet posted by a user. The high-level generative process is as follows:

(1) For each tweet d of a given user u in day s, a personal region r is drawn based on the day (**Intuitions 1** and **2**), and then draw time t based on the time Gaussian distribution of region r on the day s.
(2) A topic z is drawn based on user u's topic preference and the sampled region r (**Intuition 3**).
(3) The location ℓ and each word w are drawn based on the location and word distributions of the topic z and region r (**Intuitions 4** and **5**).

Next, we introduce the details of the generative process of each step.

In EW^4, we employ Chinese Restaurant Process (CRP) to draw the region. CRP is a stochastic process in which customers select seats at a restaurant with an infinite number of tables. The first customer randomly selects a table to sit, while the other customers can either sit at an occupied table i with probability of $\frac{n_i}{n+\beta}$, or sit at a new table with probability of $\frac{\beta}{n+\beta}$, where n_i is the number of customers at table i, and n is the total number of customers in the restaurant. As a Bayesian nonparametric approach, CRP is effective in clustering data (*i.e.,* customers) into clusters (*i.e.,* tables), and it can automatically estimate how many clusters are needed to model the data.

In our problem setting, given the historical locations that have been visited by a user, CRP automatically discovers personal regions for this user. More specifically, given the day s, user u can either select an existing region ($r \in R_u$), or create and select a new region ($r \notin R_u$) . The probability that u selects a region r is defined as follows:

$$CRP(r|u,s) = \begin{cases} \dfrac{\beta}{\sum_{r'=1}^{|R_u|} n_{s,r'}^{SR} + \beta} & r \notin R_u \\ \dfrac{n_{s,r}^{SR}}{\sum_{r'=1}^{|R_u|} n_{s,r'}^{SR} + \beta} & r \in R_u \end{cases} \tag{1}$$

Based on the sampled region r, the time t is drawn based on Gaussian distribution $\mathcal{N}(td(t,\nu_{r,s})|\nu_{r,s},\lambda_{r,s}^{-1})$ (**Intuition 2**). Note that we do not sample time t based on its exact time point, *e.g.,* 11:00 pm. Instead, the time t is generated based on the time difference $td(t,\nu_{r,s})$ between it and the mean time of the time Gaussian of region r on day s, *i.e.,* $t \sim \mathcal{N}(td(t,\nu_{r,s})|\nu_{r,s},\sigma_{r,s})$. In other words, the time that is close to the mean time is more likely to be sampled. Since the time in a day is cyclic, the time difference is always less than or equal to 12 hours, *e.g.,* the time difference between 11:00 pm and 1:00 am is 2 hours.

Parameterized by the topic preference of the user u and the sampled region r, a topic z is drawn from the region-specific probability measure G_r over topic space, where the multinomial mixing proportions of G_r is denoted by $\boldsymbol{\theta}_r$ (**Intuition 3**).

After selecting the region and topic, we draw the location ℓ and each word w. As stated in **Intuition 4**, 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). Here we use a switch c^L to decide which one accounts for the location selection: if $c^L = 1$, the location is sampled based on the topic-specific multinomial distribution over locations $\boldsymbol{\phi}_z^{ZL}$, and if $c^L = 0$, the location is sampled based on the Gaussian distribution $\mathcal{N}(\ell|\boldsymbol{\mu}_r,\boldsymbol{\Lambda}_r^{-1})$ of region r.

Since word selection is also influenced by both region and topic (**Intuition 5**), we introduce another switch c^W. If $c^W = 1$, the word is sampled based on the topic-specific multinomial distribu-

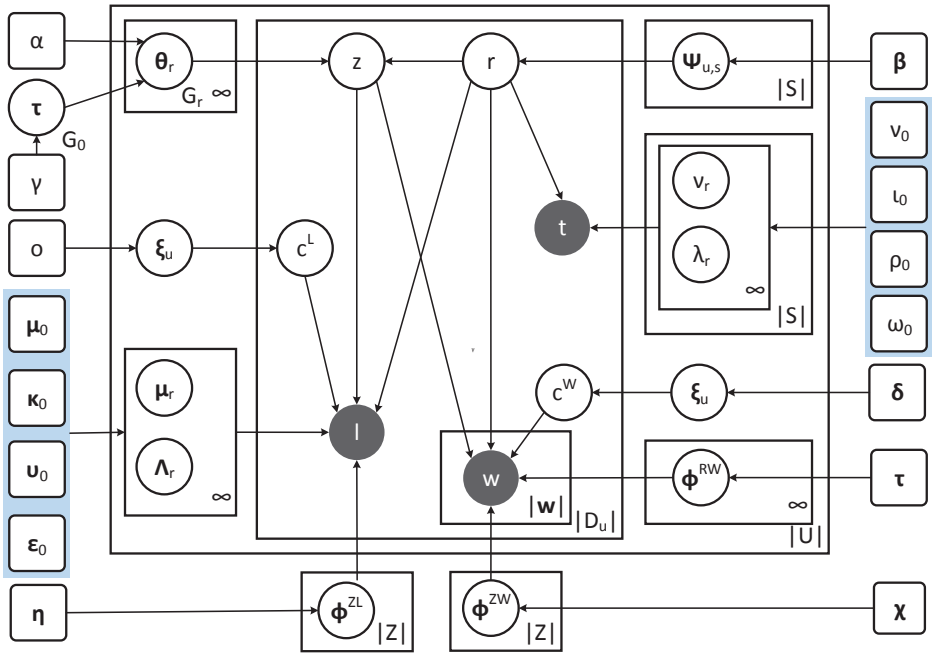


Fig. 5. The graphical representation of proposed model EW^4

tion over words ϕ_z^{ZW}, and if $c^W = 0$, the word is sampled based on the region-specific multinomial distribution over words ϕ_r^{RW}. Since a tweet is very short (limited within 140 characters), we assume all the words in one tweet come from the same topic.

Note that unlike W^4, we assume different users will make different trade-offs between topics and regions when selecting locations and words. Thus, c^L and c^W in EW^4 are drawn from two user-specific Bernoulli distributions $\boldsymbol{\xi}_u^L$ and $\boldsymbol{\xi}_u^W$, respectively.

How to set the number of topics $|Z|$ is an important issue. Most previous studies are built on topic models such as pLSA and LDA, in which the number of topics $|Z|$ needs to be empirically set. Unfortunately, it is quite hard, if possible, to tell how many topics exist in the corpus. To address this problem, we employ hierarchial Dirichlet process (HDP) in our model, which can automatically learn $|Z|$ from the data. Specifically, we introduce a global probability measure G_0 over the region-specific measure G_r, where the mixing proportion of G_0 is denoted by $\boldsymbol{\tau}$. In a finite model, the number of topics $|Z|$ is a positive integer, and $\boldsymbol{\tau}$ is drawn from the Dirichlet distribution $Dir(\gamma/|Z|, ..., \gamma/|Z|)$. After that, each $\boldsymbol{\theta}_r$ is drawn from the Dirichlet distribution $Dir(\alpha\boldsymbol{\tau})$, where α is a concentration parameter that controls the variance of the draws around $\boldsymbol{\tau}$. Taking $|Z| \to \infty$, the global topic distribution $\boldsymbol{\tau} \sim Dir(\gamma/|Z|)$, and we have $G_r \sim DP(\alpha, G_0)$, a Dirichlet process with base measure G_0 and concentration parameter α. Finally, the finite model becomes an HDP. More details about the HDP model can be found in [Teh et al. 2006].

Based on the aforementioned intuitions and notations, EW^4 generates the day, time, words, and location for each tweet posted by a user in an integrated manner. The generative process is described in Algorithm 1, and the graphical model is shown in Figure 5.

Note that we can give hyper priors for the hyper-parameters in our EW^4 model, and sample these hyper-parameters during Gibbs sampling. For example, we can give Gamma priors for α, κ_0, υ_0, *etc.*, and give Gaussian priors for $\boldsymbol{\mu_0}$ and ν_0 [Rasmussen 1999; Teh et al. 2006]. However, these hyper priors will make the model much more complicated and also slow the parameter estimation process. Thus, in this article, we empirically set these hyper-parameters at fixed values.

Algorithm 1: Generative Process of EW^4

for *each user u, (*$u = 1, ..., |U|$*)* **do**
  Draw location switch Bernoulli distribution $\boldsymbol{\xi}_u^L \sim Beta(o)$;
  Draw word switch Bernoulli distribution $\boldsymbol{\xi}_u^W \sim Beta(\delta)$;
end
Draw global topic multinomial distribution $\boldsymbol{\tau} \sim Dir(\gamma/|Z|)$;
for *each user u, (*$u = 1, ..., |U|$*)* **do**
  for *the tweet* $d_i \in D_u$ **do**
    Draw a region r based on $CRP(r|u, s_i)$, where s_i is the day of d_i;
    if $r \notin R_u$ **then**
      for *each day* $s \in S$ **do**
        Draw time distribution $\mathcal{N}(\nu_{r,s}, \lambda_{r,s}^{-1}) \sim Normal - Gamma(\nu_0, \iota_0, \rho_0, \omega_0)$;
      end
      Draw geographical distribution
      $\mathcal{N}(\boldsymbol{\mu}_r, \boldsymbol{\Lambda}_r^{-1}) \sim Normal - Wishart(\boldsymbol{\mu}_0, \kappa_0, \boldsymbol{\upsilon}_0, \epsilon_0)$;
      Draw region-specific topic multinomial distribution $\boldsymbol{\theta}_r \sim Dir(\alpha\boldsymbol{\tau})$;
      Draw region-specific word distribution $\boldsymbol{\phi}_r^{RW} \sim Dir(\zeta)$;
      Add r into R_u;
    end
    Draw a topic $z \sim \boldsymbol{\theta}_r$;
    if $z \notin Z$ **then**
      Draw topic-specific location multinomial distribution $\boldsymbol{\phi}_z^{ZL} \sim Dir(\eta)$;
      Draw topic-specific multinomial distribution $\boldsymbol{\phi}_z^{ZW} \sim Dir(\chi)$;
      Add z into Z;
    end
    Draw a time $t \sim \mathcal{N}(td(t, \nu_{r,s_i})|\nu_{r,s_i}, \lambda_{r,s_i}^{-1})$;
    Draw a location switch $c^L \sim \boldsymbol{\xi}_u^L$;
    if $c^L = 0$ **then**
      Draw a location $\ell \sim \mathcal{N}(\boldsymbol{\mu}_r, \boldsymbol{\Lambda}_r^{-1})$;
    end
    else
      Draw a location $\ell \sim \boldsymbol{\phi}_z^{ZL}$;
    end
    Draw a word switch $c^W \sim \boldsymbol{\xi}_u^W$;
    if $c^W = 0$ **then**
      for *the* $k - th$ *word (*$k = 1, ..., |\boldsymbol{w_i}|$*)* **do**
        Draw a word $w \sim \boldsymbol{\phi}_r^{RW}$;
      end
    end
    else
      for *the* $k - th$ *word (*$k = 1, ..., |\boldsymbol{w_i}|$*)* **do**
        Draw a word $w \sim \boldsymbol{\phi}_z^{ZW}$;
      end
    end
  end
end

5. PARAMETER ESTIMATION

We first introduce the sampling algorithm for parameter estimation in Section 5.1, and then discuss the time and space complexity of the sampling algorithm in Section 5.2.

5.1. Sampling Algorithm

We employ collapsed Gibbs sampling to obtain samples of the hidden variable assignments, and to estimate the unknown parameters $\{\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\phi}^{ZL}, \boldsymbol{\phi}^{ZW}, \boldsymbol{\phi}^{RW}, \boldsymbol{\xi}^{L}, \boldsymbol{\xi}^{W}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\nu}, \boldsymbol{\lambda}\}$. There are four latent variables in the model, namely, region r, topic z, the switch for location sampling c^L, and the switch for word sampling c^W.

We initialize z, c^L and c^W for each tweet by random values. Because the personal regions of each user is generated based on CRP, we create a region for each user at the initialization step, and assign all the user's tweets to the region. Then, we use two-step Gibbs sampling to obtain the samples: region r_i and topic z_i of each tweet d_i are sampled in the first step, and the two switches c_i^L and c_i^W of each tweet d_i are sampled in the second step. For each set of variables, (*e.g.,* r_i and z_i), a Gibbs sampler computes the full conditional probability for their assignments conditioned on all the other assignments (*e.g.,* $\boldsymbol{r}_{\neg i}$, $\boldsymbol{z}_{\neg i}$), while the assignments of the other set of variables (*e.g.,* $\boldsymbol{c^L}$, $\boldsymbol{c^W}$) are fixed.

For the first-step sampling, we derive the updating equation for region r_i and topic z_i for tweet d_i based on the following equation:

$$P(r_i = r, z_i = z | \boldsymbol{r}_{\neg i}, \boldsymbol{z}_{\neg i}, .) \propto \frac{P(\boldsymbol{r}, \boldsymbol{z}, .)}{P(\boldsymbol{r}_{\neg i}, \boldsymbol{z}_{\neg i}, .)}, \tag{2}$$

where other parameters involved in sampling are omitted in this equation.

However, with different c^L and c^W assignments, the generative processes of location and words of a tweet are different, which makes it difficult to get an updating equation applicable to all tweets. To solve this problem, we divide the tweet collections D into four subsets based on the assignments of c^L and c^W, namely, $D_{1,1}$, $D_{1,0}$, $D_{0,1}$ and $D_{0,0}$, where D_{c_1,c_2} denotes the collection of tweets with $c^L = c_1$ and $c^W = c_2$. Comparing to that for D, it is much easier to obtain the updating equation for tweets within each subset, because given a subset, the generative processes of locations and words of its tweets are fixed. Then, we compute the conditional probability for each set.

We first focus on $D_{1,1}$, in which tweets' locations and words are sampled according to the topic-specific distributions over locations and words, respectively. The sampling equation 2 for r_i and z_i of tweet $d_i \in D_{1,1}$ becomes:

— If $r \notin R_{u_i}$, then

$$\begin{aligned}\frac{P(\boldsymbol{r}, \boldsymbol{z}, .)}{P(\boldsymbol{r}_{\neg i}, \boldsymbol{z}_{\neg i}, .)} = & \frac{\beta}{\sum_{r'=1}^{|R_{u_i}|} n^{SR}_{s_i,r',\neg i} + \beta} \cdot \frac{\alpha\tau_z}{\sum_{z'=1}^{|Z|} n^{RZ}_{r,z',\neg i} + \alpha} \cdot \frac{n^{ZL}_{z,\ell_i,\neg_i} + \eta}{\sum_{\ell'=1}^{|L|} n^{ZL}_{z,\ell',\neg_i} + |L|\eta} \cdot \\ & \mathcal{N}(td(t_i, \nu_0) | \nu_0, \lambda_0^{-1}) \cdot \frac{\prod_{w=1}^{|V|} \prod_{y=0}^{c_{w,\boldsymbol{w}_i}} (n^{ZW}_{z,w,\neg i} + \chi + y)}{\prod_{y=0}^{c_{(.),\boldsymbol{w}_i}} \sum_{w=1}^{|V|} (n^{ZW}_{z,w,\neg i} + |V|\chi + y)}\end{aligned} \tag{3}$$

— If $r \in R_{u_i}$, then

$$\begin{aligned}\frac{P(\boldsymbol{r}, \boldsymbol{z}, .)}{P(\boldsymbol{r}_{\neg i}, \boldsymbol{z}_{\neg i}, .)} = & \frac{n^{SR}_{s_i,r,\neg i}}{\sum_{r'=1}^{|R_{u_i}|} n^{SR}_{s_i,r',\neg i} + \beta} \cdot \frac{n^{RZ}_{r,z,\neg i} + \alpha\tau_z}{\sum_{z'=1}^{|Z|} n^{RZ}_{r,z',\neg i} + \alpha} \cdot \frac{n^{ZL}_{z,\ell_i,\neg_i} + \eta}{\sum_{\ell'=1}^{|L|} n^{ZL}_{z,\ell',\neg_i} + |L|\eta} \cdot \\ & \mathcal{N}(td(t_i, \nu_r) | \nu_r, \lambda_r^{-1}) \cdot \frac{\prod_{w=1}^{|V|} \prod_{y=0}^{c_{w,\boldsymbol{w}_i}} (n^{ZW}_{z,w,\neg i} + \chi + y)}{\prod_{y=0}^{c_{(.),\boldsymbol{w}_i}} \sum_{w=1}^{|V|} (n^{ZW}_{z,w,\neg i} + |V|\chi + y)}\end{aligned} \tag{4}$$

where ℓ_i, t_i and $\boldsymbol{w}_i$ are the location, time, and words of tweet d_i; $c_{w,\boldsymbol{w}_i}$ is the number of occurrences of word w in $\boldsymbol{w}_i$, and $c_{(\cdot),\boldsymbol{w}_i}$ is the length of $\boldsymbol{w}_i$. If z is a new topic, *i.e.,* $z \notin Z$, we have $\forall \ell\ n^{ZL}_{z,\ell} = 0$, $\forall w\ n^{ZW}_{z,w} = 0$, and $\forall r\ n^{RZ}_{r,z} = 0$. $\mathcal{N}(td(t,\nu_r)|\nu_r, \lambda_r^{-1})$ is the likelihood that the temporal Gaussian distribution of r generates time t.

We estimate the parameters ν_r, λ_r for the temporal Gaussian distribution based on the time of the tweets assigned to region r, where the time collection is denoted by $\{t\}_r$. The posterior of ν_r, λ_r can be derived as follows:

$$\begin{aligned} P(\nu_r, \lambda_r|\{t\}_r, .) &\propto P(\{t\}_r|\nu_r, \lambda_r) \cdot \mathcal{NG}(\nu_r, \lambda_r|\nu_0, \iota_0, \rho_0, \lambda_0) \\ &= \prod_{t\in\{t\}_r} \mathcal{N}(td(t,\nu_r)|\nu_r, \lambda_r^{-1}) \cdot \mathcal{NG}(\nu_r, \lambda_r|\nu_0, \iota_0, \rho_0, \lambda_0) \\ &= \mathcal{NG}(\nu_r, \lambda_r|\nu'_r, \iota'_r, \rho'_r, \lambda'_r), \end{aligned} \tag{5}$$

where $\mathcal{NG}(\cdot)$ is Normal-Gamma function, and the parameters ν'_r, ι'_r, ρ'_r, λ'_r are estimated as follows:

$$\begin{aligned} \nu'_r &= \frac{\iota_0\nu_0 + |\{t\}_r|\bar{t}_r}{\iota_0 + |\{t\}_r|} \\ \iota'_r &= \iota_0 + |\{t\}_r| \\ \rho'_r &= \rho_0 + \frac{|\{t\}_r|}{2} \\ \omega'_r &= \omega_0 + \frac{1}{2}\sum_{t_k\in T_r} td(t_k - \bar{t}_r)^2 + \frac{\iota_0|\{t\}_r| \cdot td(\bar{t}_r - \nu_0)^2}{2(\iota_0 + |\{t\}_r|)} \end{aligned} \tag{6}$$

In the above equations, $\bar{t}_r$ is the average time of tweets in region r. Given Equations 5 and 6, we can update ν_r, λ_r as follows:

$$\begin{aligned} \nu_r &= \nu'_r \\ \lambda_r &= \frac{\rho'_r}{\omega'_r} \end{aligned} \tag{7}$$

The details of the equations about Gaussian parameters can be found in [Rasmussen 1999; Murphy 2007].

Note that given a collection of time, we can get two Gaussian distributions with different ν_r and λ_r, and the ν_r of the two distributions are 12-hour apart from each other. For example, ν_r for time 1:00 and 23:00 can be either 0:00 or 12:00. Obviously, 0:00 is a better choice for the mean, since it is closer to 1:00 and 23:00 comparing with 12:00. As a result, the value of λ_r for $\nu_r = 0:00$ is larger. Thus, between the two sets of ν_r and λ_r, we choose the ν_r, λ_r pair with the greater λ_r value as the mean and precision for the temporal Gaussian distribution.

For tweet d_i in the subset $D_{1,0}$, the Equation 2 for sampling the region r_i and topic z_i is as follows:

— If $r \notin R_{u_i}$, then

$$\begin{aligned} \frac{P(\boldsymbol{r}, \boldsymbol{z}, .)}{P(\boldsymbol{r}_{\neg i}, \boldsymbol{z}_{\neg i}, .)} &= \frac{\beta}{\sum_{r'=1}^{|R_{u_i}|} n^{SR}_{s_i,r',\neg i} + \beta} \cdot \frac{\alpha\tau_z}{\sum_{z'=1}^{|Z|} n^{RZ}_{r,z',\neg i} + \alpha} \cdot \frac{n^{ZL}_{z,\ell_i,\neg_i} + \eta}{\sum_{\ell'=1}^{|L|} n^{ZL}_{z,\ell',\neg_i} + |L|\eta} \cdot \\ &\quad \mathcal{N}(td(t_i,\nu_0)|\nu_0, \lambda_0^{-1}) \cdot \frac{\prod_{w=1}^{|V|}\prod_{y=0}^{c_{w,\boldsymbol{w}_i}}(\zeta + y)}{\prod_{y=0}^{c_{(\cdot),\boldsymbol{w}_i}}\sum_{w=1}^{|V|}(|V|\zeta + y)} \end{aligned} \tag{8}$$

— If $r \in R_{u_i}$, then

$$\frac{P(\boldsymbol{r},\boldsymbol{z},.)}{P(\boldsymbol{r}_{\neg i},\boldsymbol{z}_{\neg i},.)} = \frac{n^{SR}_{s_i,r,\neg i}}{\sum_{r'=1}^{|R_{u_i}|} n^{SR}_{s_i,r',\neg i}+\beta} \cdot \frac{n^{RZ}_{r,z,\neg i}+\alpha\tau_z}{\sum_{z'=1}^{|Z|} n^{RZ}_{r,z',\neg i}+\alpha} \cdot \frac{n^{ZL}_{z,\ell_i,\neg_i}+\eta}{\sum_{\ell'=1}^{|L|} n^{ZL}_{z,\ell',\neg_i}+|L|\eta} \cdot \mathcal{N}(td(t_i,\nu_r)|\nu_r,\lambda_r^{-1}) \cdot \frac{\prod_{w=1}^{|V|}\prod_{y=0}^{c_{w,\boldsymbol{w}_i}}(n^{RW}_{r,w,\neg i}+\zeta+y)}{\prod_{y=0}^{c_{(.),\boldsymbol{w}_i}}\sum_{w=1}^{|V|}(n^{RW}_{r,w,\neg i}+|V|\zeta+y)} \quad (9)$$

For tweet d_i in subset $D_{0,1}$, the Equation 2 for sampling the region r_i and topic z_i is as follows:

— If $r \notin R_{u_i}$, then

$$\frac{P(\boldsymbol{r},\boldsymbol{z},.)}{P(\boldsymbol{r}_{\neg i},\boldsymbol{z}_{\neg i},.)} = \frac{\beta}{\sum_{r'=1}^{|R_{u_i}|} n^{SR}_{s_i,r',\neg i}+\beta} \cdot \frac{\alpha\tau_z}{\sum_{z'=1}^{|Z|} n^{RZ}_{r,z',\neg i}+\alpha} \cdot \mathcal{N}(\ell_i|\boldsymbol{\mu}_0,\boldsymbol{\Lambda}_0^{-1}) \cdot \mathcal{N}(td(t_i,\nu_0)|\nu_0,\lambda_0^{-1}) \cdot \frac{\prod_{w=1}^{|V|}\prod_{y=0}^{c_{w,\boldsymbol{w}_i}}(n^{ZW}_{z,w,\neg i}+\chi+y)}{\prod_{y=0}^{c_{(.),\boldsymbol{w}_i}}\sum_{w=1}^{|V|}(n^{ZW}_{z,w,\neg i}+|V|\chi+y)} \quad (10)$$

— If $r \in R_{u_i}$, then

$$\frac{P(\boldsymbol{r},\boldsymbol{z},.)}{P(\boldsymbol{r}_{\neg i},\boldsymbol{z}_{\neg i},.)} = \frac{n^{SR}_{s_i,r,\neg i}}{\sum_{r'=1}^{|R_{u_i}|} n^{SR}_{s_i,r',\neg i}+\beta} \cdot \frac{n^{RZ}_{r,z,\neg i}+\alpha\tau_z}{\sum_{z'=1}^{|Z|} n^{RZ}_{r,z',\neg i}+\alpha} \cdot \mathcal{N}(\ell_i|\boldsymbol{\mu}_r,\boldsymbol{\Lambda}_r^{-1}) \cdot \mathcal{N}(td(t_i,\nu_r)|\nu_r,\lambda_r^{-1}) \cdot \frac{\prod_{w=1}^{|V|}\prod_{y=0}^{c_{w,\boldsymbol{w}_i}}(n^{ZW}_{z,w,\neg i}+\chi+y)}{\prod_{y=0}^{c_{(.),\boldsymbol{w}_i}}\sum_{w=1}^{|V|}(n^{ZW}_{z,w,\neg i}+|V|\chi+y)} \quad (11)$$

The parameters $\boldsymbol{\mu}_r$ and $\boldsymbol{\Lambda}_r$ for the spatial Gaussian distribution of region r can be estimated based on the coordinates of tweet locations assigned to r with $c^L = 0$. We use $\{\boldsymbol{cd}\}_r$ to denote the collection of such coordinates, and obtain the posterior of $\boldsymbol{\mu}_r$ and $\boldsymbol{\Lambda}_r$ as follows:

$$\begin{aligned} P(\boldsymbol{\mu}_r,\boldsymbol{\Lambda}_r|\{\boldsymbol{cd}\}_r,.) &\propto P(\{\boldsymbol{cd}\}_r|\boldsymbol{\mu}_r,\boldsymbol{\Lambda}_r)\cdot\mathcal{NW}(\boldsymbol{\mu}_r,\boldsymbol{\Lambda}_r|\boldsymbol{\mu_0},\kappa_0,\upsilon_0,\boldsymbol{\epsilon_0}) \\ &= \prod_{\boldsymbol{cd}\in\{\boldsymbol{cd}\}_r}\mathcal{N}(\boldsymbol{cd}|\boldsymbol{\mu}_r,\boldsymbol{\Lambda}_r^{-1})\cdot\mathcal{NW}(\boldsymbol{\mu}_r,\boldsymbol{\Lambda}_r|\boldsymbol{\mu_0},\kappa_0,\upsilon_0,\boldsymbol{\epsilon_0}) \\ &= \mathcal{NW}(\boldsymbol{\mu_r}',\kappa_r',\upsilon_r',\boldsymbol{\epsilon_r'}), \end{aligned} \quad (12)$$

where $\mathcal{NW}(\cdot)$ is Normal-Wishart function, and $\boldsymbol{\mu_r}'$, κ_r', υ_r', $\boldsymbol{\epsilon_r'}$ are estimated as follows:

$$\begin{aligned} \boldsymbol{\mu}_r' &= \frac{\kappa_0\boldsymbol{\mu}_0+|\{\boldsymbol{cd}\}_r|\overline{\boldsymbol{cd}}_r}{\kappa_0+|\{\boldsymbol{cd}\}_r|} \\ \kappa_r' &= \kappa_0+|\{\boldsymbol{cd}\}_r| \\ \upsilon_r' &= \upsilon_0+|\{\boldsymbol{cd}\}_r| \\ \boldsymbol{\epsilon}_r' &= \boldsymbol{\epsilon}_0+\sum_{\boldsymbol{cd}\in\{\boldsymbol{cd}\}_r}(\boldsymbol{cd}-\overline{\boldsymbol{cd}}_r)(\boldsymbol{cd}-\overline{\boldsymbol{cd}}_r)^T+\frac{\kappa_0|\{\boldsymbol{cd}\}_r|}{\kappa_0+|\{\boldsymbol{cd}\}_r|}(\boldsymbol{\mu}_0-\overline{\boldsymbol{cd}}_r)(\boldsymbol{\mu}_0-\overline{\boldsymbol{cd}}_r)^T \end{aligned} \quad (13)$$

In the above equation, $\overline{\boldsymbol{cd}}_r$ is the mean of $\{\boldsymbol{cd}\}_r$. Given Equations 12 and 13, we can update $\boldsymbol{\mu}_r$, $\boldsymbol{\Lambda}_r$ as follows:

$$\begin{aligned} \boldsymbol{\mu}_r &= \boldsymbol{\mu}_r' \\ \boldsymbol{\Lambda}_r &= \upsilon_r'\cdot\boldsymbol{\epsilon}_r'^{-1} \end{aligned} \quad (14)$$

Last, we can sample the region r_i and topic z_i for tweet d_i in subset $D_{0,0}$ (details can be found in the published version from ACM Digital Library).

When the number of topics $|Z|$ changes, and when a sampling iteration is finished, we sample new global topic distribution $\boldsymbol{\tau}$ (details can be found in the published version from ACM Digital Library).

After sampling region r_i and topic z_i for all tweets $d_i \in D$, we sample c_i^L and c_i^W (details can be found in the published version from ACM Digital Library).

After sampling a sufficient number of iterations, we calculate the parameters as follows:

$$
\begin{aligned}
\boldsymbol{\psi}_{u,s,r} &= P(r|s) = \frac{n_{s,r}^{SR} + \frac{\beta}{|R_u|}}{\sum_{r'=1}^{|R_u|} n_{s,r'}^{SR} + \beta} \\
\boldsymbol{\theta}_r &= P(z|r) = \frac{n_{r,z}^{RZ} + \alpha\tau_z}{\sum_{z'=1}^{|Z|} n_{r,z'}^{RZ} + \alpha} \\
\boldsymbol{\xi}_u^L &= P(c^L = 1|u) = \frac{n_{u,(1)}^{UCL} + o_1}{n_{u,(0)}^{UCL} + n_{u,(1)}^{UCL} + o_0 + o_1} \\
\boldsymbol{\xi}_u^W &= P(c^W = 1|u) = \frac{n_{u,(1)}^{UCW} + \delta_1}{n_{u,(0)}^{UCW} + n_{u,(1)}^{UCW} + \delta_0 + \delta_1} \\
\boldsymbol{\phi}_z^{ZL} &= P(\ell|z) = \frac{n_{z,l}^{ZL} + \eta}{\sum_{\ell'=1}^{|L|} n_{z,\ell'}^{ZL} + |L|\eta} \\
\boldsymbol{\phi}_z^{ZW} &= P(w|z) = \frac{n_{z,w}^{ZW} + \chi}{\sum_{w'=1}^{|V|} n_{z,w'}^{ZW} + |W|\chi} \\
\boldsymbol{\phi}_r^{RW} &= P(w|r) = \frac{n_{r,w}^{RW} + \zeta}{\sum_{w'=1}^{|V|} n_{r,w'}^{RW} + |W|\zeta}
\end{aligned}
$$

5.2. Time and Space Complexity

We use Gibbs sampling to estimate the parameters. For each tweet in each iteration, we need to calculate its probability distributions over topics and personal regions, and over switches c^L and c^W. The process is iterated for T times. The total time complexity is therefore $O(T|D|(|R_{max}||Z_{max}||W_{max}| + |c^L||c^W|))$, where $|R_{max}|$ is the maximum number of personal regions of users, Z_{max} is the maximum number of topics, $|W_{max}|$ is the maximum length of a tweet (*i.e.,* 140), and $|c^L| = |c^W| = 2$.

The space required by EW^4 is as follows:

— $O(|U||R_{max}|)$ and $O(|U||S||R_{max}|)$ for the parameters of users' time and spatial Gaussian distributions;
— $O(|Z_{max}|)$ for the topic proportions $\boldsymbol{\pi}$ of the global measure G_0;
— $O(4 \times |D|)$ for the topic, region, c^L, c^W assignments of documents;
— $O(|Z_{max}|(|L| + |W|))$ for the count of locations' and words' topic assignments;
— $O(|U|(|c^L| + |c^W|))$ for the number of c^L and c^W selections of each user;
— $O(|U||R_{max}|(|Z_{max}| + |W|))$ for the topic and word counts of each personal region;

Aggregating them together, the total space complexity is $O((|U| + |D|) + |Z_{max}|(|L| + |W|) + |U||R_{max}|(|Z_{max}| + |W|))$. Since the matrices for locations and words are very sparse, thus the storage can be significantly reduced by utilizing sparse matrices. When reading out parameters, we need $O(|U||R_{max}|(|Z_{max}|+|W|)+|Z_{max}|(|L|+|W|)+|U|)$ to keep the parameters (See Table I).

For the two data sets (WW and USA, See Section 7.1), the memory required is less than 1 GB. As future work, it would be interesting to improve the complexity to handle data of large scale [Cui et al. 2014].

6. APPLICATIONS

The proposed model EW^4 involves four aspects of user's mobility behavior (*i.e.,* who, where, when, and what). As the online component of our framework, the model can infer missing information in some of four aspects given information available from other aspects. A variety of applications can be built on top of the model and we name a few as examples.

Requirement-aware location recommendation. Location recommendation aims to recommend new locations for users to visit. Sometimes a user may have clear requirement about the recommendation, *e.g.,* a user wants to have pizza at 7:00 PM. Obviously, the requirement explicitly reveals a user' preference, and thus can be utilized for making recommendations. However, to the best of our knowledge, none of previous studies on POI recommendation has considered users' requirements. Since EW^4 jointly models the who, where, when and what factors, it is able to utilize both time and need (in the form of short text or keywords) to make requirement-aware recommendations. Formally, given a user u, day s, time t and a set of words $\boldsymbol{w}$ that describe user need, the candidate locations are ranked by:

$$
\begin{aligned}
P(\ell|u,s,t,\boldsymbol{w}) &= \frac{\sum_{z=1}^{|Z|}\sum_{r=1}^{|R_u|} P(u,s,t,r,z,\boldsymbol{w},\ell)}{\sum_{z=1}^{|Z|}\sum_{r=1}^{|R_u|}\sum_{\ell'=1}^{|L|} P(u,s,t,r,z,\boldsymbol{w},\ell')} \\
&\propto \sum_{z=1}^{|Z|}\sum_{r=1}^{|R_u|} P(r|u,s)P(z|r)P(t|r)P(l|r,z)P(\boldsymbol{w}|r,z) \\
&= \sum_{z=1}^{|Z|}\sum_{r=1}^{|R_u|} \psi_{u,s,r}\cdot\theta_{r,z}\cdot\mathcal{N}(td(t,\nu_r)|\nu_r,\lambda_r^{-1})\cdot(\xi_{u,0}^{L}\cdot\mathcal{N}(\ell|\boldsymbol{\mu}_r,\boldsymbol{\Lambda}_r^{-1})+\xi_{u,1}^{L}\cdot\phi_{z,\ell}^{ZL})\cdot \\
&\quad \prod_{w\in\boldsymbol{w}}(\xi_{u,0}^{W}\cdot\phi_{r,w}^{RW}+\xi_{u,1}^{W}\cdot\phi_{z,w}^{ZW}) \qquad (15)
\end{aligned}
$$

Location prediction for tweet. Given a tweet with its content in words, user, 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 [Cho et al. 2011; Cheng et al. 2011] that geographical locations can be used to predict user's behavior, discover users' interests, and deliver location-based advertisement or content. However, not all tweets are explicitly annotated with geographical locations. Hence, location prediction for tweets is a very important application and can be used to facilitate many applications.

A number of methods have been proposed for this task [Li et al. 2011; Kinsella et al. 2011; Eisenstein et al. 2010; Wing and Baldridge 2011; Hong et al. 2012; Ahmed et al. 2013]. The studies [Li et al. 2011; Kinsella et al. 2011] 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. The work [Wing and Baldridge 2011] segments the world into grids, and employs supervised models to predict a grid for a given text. The recent proposals [Hong et al. 2012; Ahmed et al. 2013] present approaches for predicting geographic coordinates of a text from a user. Since EW^4 can make both kinds of predictions for a text from a user, namely, predicting location identifiers and geographic coordinates. Our method is also able to take the time factor into consideration.

Formally, given a user u, day s, time t, and words $\boldsymbol{w}$, a location ℓ is predicted based on Equation 15. Then, the top 1 location is returned as the prediction result.

Activity prediction. EW^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:

$$\begin{aligned} P(w|u,s,t) &= \frac{\sum_{z=1}^{|Z|}\sum_{r=1}^{|R_u|} P(u,s,t,r,z,w)}{\sum_{z=1}^{|Z|}\sum_{r=1}^{|R_u|}\sum_{w'=1}^{|V|} P(u,s,t,r,z,w')} \\ &\propto \sum_{z=1}^{|Z|}\sum_{r=1}^{|R_u|} P(r|u,s)P(z|r)P(t|r)P(w|r,z) \\ &= \sum_{z=1}^{|Z|}\sum_{r=1}^{|R_u|} \psi_{u,s,r}\cdot\theta_{r,z}\cdot\mathcal{N}(td(t,\nu_r)|\nu_r,\lambda_r^{-1})\cdot(\xi_{u,0}^{W}\cdot\phi_{r,w}^{RW}+\xi_{u,1}^{W}\cdot\phi_{z,w}^{ZW}) \end{aligned} \quad (16)$$

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:

$$\begin{aligned} P(u|\ell,s,t) &= \frac{\sum_{z=1}^{|Z|}\sum_{r=1}^{|R_u|} P(u,s,t,r,z,\ell)}{\sum_{u'=1}^{|U|}\sum_{z=1}^{|Z|}\sum_{r=1}^{|R_{u'}|} P(u',s,t,r,z,\ell)} \\ &\propto \sum_{z=1}^{|Z|}\sum_{r=1}^{|R_u|} P(r|u,s)P(z|r)P(t|r)P(l|r,z) \\ &= \sum_{z=1}^{|Z|}\sum_{r=1}^{|R_u|} \psi_{u,s,r}\cdot\theta_{r,z}\cdot\mathcal{N}(td(t,\nu_r)|\nu_r,\lambda_r^{-1})\cdot \\ &\quad (\xi_{u,0}^{L}\cdot\mathcal{N}(\ell|\boldsymbol{\mu}_r,\boldsymbol{\Lambda}_r^{-1})+\xi_{u,1}^{L}\cdot\phi_{z,\ell}^{ZL}) \end{aligned} \quad (17)$$

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:

$$\begin{aligned} P(\ell|u,s,t) &= \frac{\sum_{z=1}^{|Z|}\sum_{r=1}^{|R_u|} P(u,s,t,r,z,\ell)}{\sum_{z=1}^{|Z|}\sum_{r=1}^{|R_u|}\sum_{\ell'=1}^{|L|} P(u,s,t,r,z,\ell')} \\ &\propto \sum_{z=1}^{|Z|}\sum_{r=1}^{|R_u|} P(r|u,s)P(z|r)P(t|r)P(l|r,z) \\ &= \sum_{z=1}^{|Z|}\sum_{r=1}^{|R_u|} \psi_{u,s,r}\cdot\theta_{r,z}\cdot\mathcal{N}(td(t,\nu_r)|\nu_r,\lambda_r^{-1})\cdot \\ &\quad (\xi_{u,0}^{L}\cdot\mathcal{N}(\ell|\boldsymbol{\mu}_r,\boldsymbol{\Lambda}_r^{-1})+\xi_{u,1}^{L}\cdot\phi_{z,\ell}^{ZL}) \end{aligned} \quad (18)$$

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 $P(\boldsymbol{w}|u,s,t,\ell)$, where $\boldsymbol{w}$ is the word vector of a

candidate tweets.

$$
\begin{aligned}
P(\boldsymbol{w}|u,s,t,\ell) &= \frac{\sum_{z=1}^{|Z|}\sum_{r=1}^{|R_u|} P(u,s,t,r,z,\boldsymbol{w},\ell)}{\sum_{z=1}^{|Z|}\sum_{r=1}^{|R_u|}\sum_{\boldsymbol{w}} P(u,s,t,r,z,\boldsymbol{w}',\ell)} \\
&\propto \sum_{z=1}^{|Z|}\sum_{r=1}^{|R_u|} P(r|u,s)P(z|r)P(t|r)P(l|r,z)P(\boldsymbol{w}|r,z) \\
&= \sum_{z=1}^{|Z|}\sum_{r=1}^{|R_u|} \psi_{u,s,r}\cdot\theta_{r,z}\cdot\mathcal{N}(td(t,\nu_r)|\nu_r,\lambda_r^{-1})\cdot(\xi_{u,0}^{L}\cdot\mathcal{N}(\ell|\boldsymbol{\mu}_r,\boldsymbol{\Lambda}_r^{-1})+\xi_{u,1}^{L}\cdot\phi_{z,\ell}^{ZL})\cdot \\
&\quad \prod_{w\in\boldsymbol{w}}(\xi_{u,0}^{W}\cdot\phi_{r,w}^{RW}+\xi_{u,1}^{W}\cdot\phi_{z,w}^{ZW}) \qquad (19)
\end{aligned}
$$

7. EXPERIMENTS

We evaluate the proposed model in this section. We first evaluate the accuracy of EW[4] for the application of *location prediction for tweets* and *requirement-aware location recommendation* in Sections 7.2 and 7.3 against several state-of-the-art baselines. Then, we present samples of the discovered topics and the mobility patterns of users in Section 7.4. Results of other example applications of the proposed model are reported in Section 7.5.

7.1. Data set

Two real-world data sets are used in the experiments, namely, *WW* data set and *USA* data set. Next, we give the details about the two data sets.

WW Data set. Foursquare users can associate their accounts to Twitter, so that when they make check-ins in Foursquare, corresponding tweets will be posted in Twitter. Using the streaming API provided by Twitter[2], we collect 3,478,394 Foursquare check-ins from November 1, 2012 to February 13, 2013, among which 1,322,437 contains shouts (short messages) written in English alphabets. We examine the users who posted English shouts, and remove the inactive users who visited fewer than 5 different locations. Since users may check in when traveling to new places, and incorporating such check-ins will make it hard to estimate personal regions. Thus, we filter out the outlier check-ins as follows: we train GMM for each user, and remove invalid Gaussian components whose weights are smaller than 0.1. Check-ins that are most close to these invalid components are deleted. We iterate this process for each user. In the end, 89,007 check-ins are left after pre-processing. We refer to this data set as *WW* (World-wide) data set as the tweets are from users in different countries.

USA data set. This data set is the GeoText[3] (Geo-tagged Microblog Corpus) published by researchers from Carnegie Mellon University [Eisenstein et al. 2010]. This data set comprises messages from geo-located microblog users approximately in the United States. Each message is associated with its geographic coordinates. 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. After that, we apply the same pre-processing with WW to this data set.

We remove stop-words from the text in both data sets. The statistics of the data sets after pre-processing is shown in Table II. For each data set, we randomly split the documents (tweets or messages) into three collections in proportion of 8:1:2 as the training set, development set, and test

[2] https://dev.twitter.com/docs/streaming-apis
[3] http://www.ark.cs.cmu.edu/GeoText/

Table II. Statistics of the two data sets

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

set, respectively. We do this by following a previous work [Cho et al. 2011] [4]. The data sets used in this paper are available online [5].

7.2. 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.

7.2.1. 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.

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.

7.2.2. Baseline Methods. We compare our model with 7 baseline methods to evaluate the performance, including the state-of-the-art models for predicting locations for text.

KL-divergence based Model (KL) [Li et al. 2011; Kinsella et al. 2011]. 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. The location with smallest KL-divergence is returned as the prediction result, and its coordinates are used to calculate the error distance.

Mean Coordinates (Mean). This model estimates the mean coordinates of visited locations for each user. Given a tweet, it returns the location that is closest to its author's mean coordinates as the prediction result.

Popular Location (Pop). This model first finds out the location for each user that she visited most frequently. Given a tweet, it returns the most frequently visited location of its author as the prediction result.

Topic+Region Model (TR) [Hong et al. 2012]. This model captures the user preference over latent regions and topics. The locations, which are treated as geographic coordinates, are generated from the Gaussian distributions of regions, and words are generated based on the topics and regions. In addition, the latent regions in this model are not personal. Given a tweet from a user, TR can predict the geographic coordinates of the tweet.

Hierarchical Geographical Model (HG) [Ahmed et al. 2013]. This model organizes geographical regions in a hierarchy, where regions in lower level are more specific w.r.t geographical area and topics. Each tweet is generated by a path from the root region to a leaf region, while the text content is drawn based on topics and the language model of the selected leaf region. Similar to that of TR, regions in HG are also global.

[4] Note that the model proposed in [Cho et al. 2011] does not contain tuning parameters, and the authors randomly split the data set into training and test sets in proportion of 8:2 but do not create a development set.

[5] http://www.ntu.edu.sg/home/gaocong/datacode.htm

Table III. Comparison of baseline methods with $\text{EW}^4_{\setminus T}$ and EW^4

Factors in modeling	KL	Mean	Pop	TR	HG	ST	W^4	$\text{EW}^4_{\setminus T}$	EW^4
Who (User)	×	√	√	√	√	√	√	√	√
Where (Geo)	×	√	×	GlbR	GlbR	GlbR	PsnR	PsnR	PsnR
When (Time)	×	×	×	×	×	×	√	×	√
What (Words)	√	×	×	√	√	√	√	√	√
Manually setting tuning parameters	×	×	×	√	×	√	√	×	×

Spatial Topic Model (ST) [Hu and Ester 2013]. In ST model, each user has a distribution over global regions and topics. Different from TR, ST considers the identifiers of locations, and each topic has a distribution over location identifiers. ST assumes each tweet has a region and a topic: the topic determines the text content of the tweet, and topic and region together influence the draw of a location, *i.e.,* the sampling probability of a location is proportional to the product of the likelihoods of the topic generating the location and the region's Gaussian distribution generating the location.

Who+Where+When+What (W^4) [Yuan et al. 2013b]. This model is built based on similar intuitions as in EW^4. Specifically, for each tweet, a personal region is first drawn based on its user and time, and then a topic is drawn based on the user's topic distribution at that region. Finally, the topic and region together generate the location and the words of the tweet. The differences between W^4 and EW^4 are summarized in Section 2.

Note that both TR and HG are designed to predict the geographical coordinates, and cannot return the location identifier. Thus we cannot compute Acc for the two baselines. In order to compare with those approaches in terms of Acc, we identify the location identifier for the predicted geographic coordinates by finding the nearest location to the coordinates.

All baselines have been optimized by the development set. Specifically, for TR, the numbers of regions for WW and USA data sets are 500 and 600, respectively. For HG, the priors over topics, topics mixing vectors, parameter λ and ω are all set to 0.1, and the numbers of regions for both data sets are 600. The number of topics for TR and HG are 50. For W^4, the number of topics for WW and USA data sets are 10 and 20, parameters λ, κ for the two data sets are both 0.6 and 0.1, respectively.

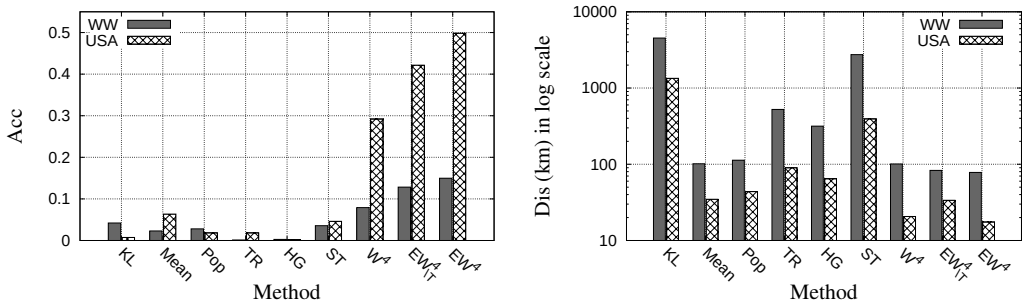
7.2.3. Our Proposed Methods. The above baseline models are compared with EW^4, the model proposed in this article.

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

Enhanced Who+Where+What ($\text{EW}^4_{\setminus T}$). Except our preliminary work [Yuan et al. 2013b], none of existing studies makes use of the time factor in prediction. To study the performance of our model without time factor, we consider a simplified version of EW^4, known as $\text{EW}^4_{\setminus T}$, which does not consider the time factor. Note that $\text{EW}^4_{\setminus T}$ exploits the similar set of aspects as the baseline approaches TR HG and ST do, but its modeling method is different from theirs.

7.2.4. Experimental results. We compare the prediction performance of the 9 methods (KL, Mean, Pop, TR, ST, HG, W^4, $\text{EW}^4_{\setminus T}$, and EW^4). The Dis and Acc of each method are reported in Figure 6. Note that only W^4 and EW^4 make use of the time information in prediction.

As shown in Figure 6, our preliminary model W^4 outperforms the state-of-the-art baseline methods KL, TR ST and HG significantly in terms of both Acc and Dis. The Acc of W^4 on *WW* and *USA* are 0.0792 and 0.2920, outperforming KL in terms of Acc by 88.50% and 3953.04% on the two data sets, respectively. The Dis of W^4 on *WW* is 100.93 km on *USA* is and 20.63 km. It reduces the Dis of TR by 80.73% and 77.02%, and reduce the Dis of HG by 68.09% and 68.02%, on the two data sets, respectively. The EW^4 proposed in this article achieves Acc of 0.1498 and 0.4986 on the two data sets, which are 89.14% and 70.75% greater than that of W^4 on *WW* and *USA* data



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

Fig. 6. Prediction Performance of all methods

sets, respectively. The Dis of EW^4 are 78.16 km on *WW* and 17.47 km on *USA*, indicating that EW^4 reduces the average error distance against W^4 by 22.55% and 15.31% on the two data sets, respectively.

Mean and Pop are two model-free baselines which do not make use of time and word information. Interestingly, they achieve comparable Acc and much better Dis against other baselines, even include the complex models TR, ST and HG. Potential reason is a user's mobility is constrained in a limited region which centers at a specific point, particularly when the user only visited very few locations. Thus, using their mean coordinates and mostly visited locations as predictions already achieves satisfactory results. However, compared with them, our proposed model EW^4 improves their Acc by more than 651.30% in Acc, and reduces their Dis by more than 30.27% on both data sets.

KL is designed to predict the location label for short text. Because it does not exploit geographic coordinates 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 models 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 II, and thus the language models of location are usually sparse, limiting the prediction performance of KL.

Different from KL, TR and HG are designed to predict the geographic coordinates for short text. They return 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 data sets. 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. Comparing with TR, HG achieves a better performance, because it exploits the hierarchical relations between regions. However, Acc of TR and HG are approximate to 0, since the means of the global regions are less likely to be the exact locations of individuals' tweets.

ST makes use of both the identifers and geographic coordinates of locations, but its Dis is the worst among the topic-model based methods, and its Acc is better than TR and HG that do not use location identifiers. We checked the results, and found ST often returns the same location for the test tweets posted by the same user. After investigation, we found that many of the returned locations lie in the centers of regions with a quite large precision value. The large likelihood that the regions' Gaussian distributions generate the location makes the location always receive the greatest ranking score among the candidate locations.

Our model $\text{EW}^4_{\setminus T}$ utilizes the same types of information as do TR, HG and ST, but it outperforms the latter three baselines significantly. The reasons are two-fold. First, the latent geographic regions

in $\text{EW}^4_{\setminus T}$ are personal while the latent geographic regions in the three baselines are global for all the users. Hence, the regions in $\text{EW}^4_{\setminus T}$ can describe individuals' mobility areas more precisely than the regions in TR, HG and ST. Second, both the location identifiers and the geographic information of locations are used by $\text{EW}^4_{\setminus T}$ to enhance the prediction, while TR and HG only exploit the geographic coordinates of locations.

EW^4 outperforms $\text{EW}^4_{\setminus T}$ in terms of both measures. This is because EW^4 incorporates the time factor in its model, which can further improve the prediction results. EW^4 is capable of capturing the user's mobility patterns in terms of geographic, temporal, and activity aspects.

Comparing with W^4, the enhanced version EW^4 achieves better results in terms of both Acc and Dis. The reasons are three-fold: 1) EW^4 is designed under the framework of HDP, which is more robust to the overfitting problem; 2) in EW^4, users can have different numbers of regions, which can be automatically learnt from the data by CRP. The user-specific region number can help better model users' mobility regions; 3) the weights of topic and region for the selections of locations and words are learnt from training data, which are user-specific.

7.3. Requirement-aware location recommendation

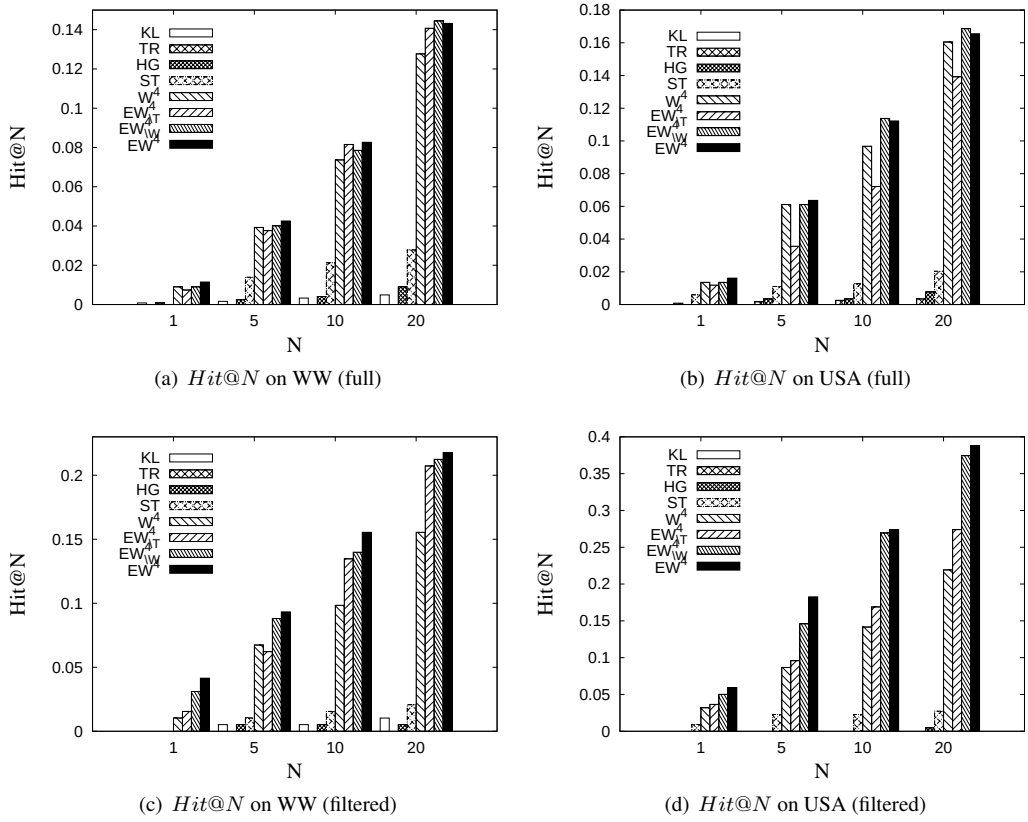
Given a user and the user's specific requirements (represented by a set of words), requirement-aware location recommendation aims to recommend a ranked list of locations that the user has not visited but might be interested in. When the target time is available, we can also incorporate the time as additional contextual information. Although requirement-aware location recommendation uses the same ranking equation as location prediction for tweets, they are two different tasks: when predicting locations for tweets, the true locations may be the locations that the users have visited many times, while for location recommendation, the true locations are new to the users, *i.e.,* the users have not visited the true locations before.

However, it is hard to evaluate the accuracy of requirement-aware location recommendation. Recall the example in Section 5 that a user wants to have pizza at 7:00 PM. To get recommendations, the user can submit a requirement-aware location recommendation query with word "pizza" and time "7:00 PM" before the target time (*e.g.,* at 1:00 PM). The only way to verify the recommendation accuracy is to check whether the user indeed visited one of the recommended locations at 7:00 PM. However, it is very difficult to collect such requirements and ground-truth recommendations for evaluating the requirement-aware location recommendation task.

In this article, we choose to use the information of a tweet (including user, time, words) in the test set as a requirement query, and return a ranked list of locations that the user has not been to (*i.e.,* has not visited in the training set). Here the time indicates the context, and the words describe the requirement of the user. In fact, we treat a location visit of a tweet query as a future event rather than a past event, and we use the location of the tweet as the ground-truth for evaluation. We admit that the tween content my not always reflect the real user requirement.

To evaluate the recommendation performance, we use the same training and development sets that are used in the task of location prediction for tweets, but only keep the tweets in test set whose locations do not appear in their users' tweets in the training set. The number of such tweets is 1,221 and 1,178 in WW and USA data sets, respectively, and they are used as a group of test data, denoted as "full". However, many of the tweets do not represent specific requirements. Thus, we ask two annotators to create another group of data sets by removing non-English (but written in English alphabets) and requirement-irrelevant tweets. Only those tweets that are validated by both of the annotators are kept. After annotation, 193 and 219 tweets are left for WW and USA data sets, respectively, as another group of test data, denoted as "filtered".

7.3.1. Evaluation Metric. We evaluate the recommendation performance of different models by Hit ratio @ N ($Hit@N$), which measures the percentage of test instances whose true locations are captured in the top N recommendations. Obviously, larger $Hit@N$ values indicate better performance. We set N to be 1, 5, 10, and 20.



(a) $Hit@N$ on WW (full)

(b) $Hit@N$ on USA (full)

(c) $Hit@N$ on WW (filtered)

(d) $Hit@N$ on USA (filtered)

Fig. 7. Recommendation Performance of all methods on both data sets

7.3.2. Methods to be evaluated. We compare the effectiveness of the methods that can utilize text for recommendation(KL, TR, ST, HG, W^4, $EW^4_{\setminus T}$, and EW^4). In order to examine the effectiveness of text, we remove the word factor from EW^4 as another baseline named $EW^4_{\setminus W}$.

7.3.3. Experimental results. The $Hit@N$ of the 8 methods are reported in Figure 7.

Among these methods, the performance of KL is the worst, because it does not exploit user, time and geographical information in recommendation. The $Hit@N$ values of TR and HG are also very low. Because no location identifiers are used in the two models, they are ineffective in recommending the unvisited location identifiers for users. Compared with TR and HG, ST achieves much better $Hit@N$ values, because it makes use of both the geographical coordinates and identifiers of locations.

Compared with KL, TR, HG, and ST, our proposed method $EW^4_{\setminus T}$ always achieves much better $Hit@N$ values on different N, even though it utilizes the same information with TR, HG, and ST. For example, before removing the requirement-irrelevant tweets, $EW^4_{\setminus T}$ outperforms the best baseline ST by 283.4% and 466.7% in terms of $Hit@10$ on WW and USA, respectively, as shown in Figures 7(a) and 7(b). After the filtering, the improvement becomes 769.0% and 640.0%, respectively, as shown in Figures 7(c) and 7(d). The improvement may be attributed to two reasons: 1) personal regions that can describe user mobility more precisely, and 2) the consideration of both identifiers and coordinates of locations. After incorporating time factor, our full model EW^4 always

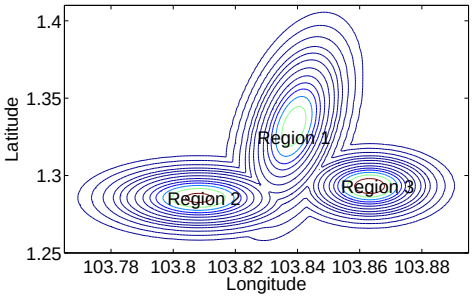


Fig. 8. Personal regions

outperforms $\text{EW}^4_{\setminus T}$, demonstrating the importance of time for requirement-aware location recommendation.

$\text{EW}^4_{\setminus W}$ is another simplified version of EW^4 that does not consider text. However, the improvement of EW^4 over $\text{EW}^4_{\setminus W}$ is not very significant, especially when N is large. For example, before removing the requirement-irrelevant tweets (full), $\text{EW}^4_{\setminus W}$ even achieves slightly better $Hit@20$ than EW^4, as shown in Figures 7(a) and 7(b). The reason would be that the noisy tweet content deteriorates the recommendation accuracy. After removing such tweets (filtered), EW^4 performs slightly better than $\text{EW}^4_{\setminus W}$ in terms of $Hit@20$. However, when N is small, the improvement of EW^4 over $\text{EW}^4_{\setminus W}$ becomes significant. For example, either before or after removing the requirement-irrelevant tweets, EW^4 always outperforms $\text{EW}^4_{\setminus W}$ by more than 19% w.r.t. $Hit@1$. The results show that text requirement is important to generate accurate recommendations among the top several results.

In summary, our full model EW^4 achieves superior accuracy in recommending locations based on the target time and specific requirements. In addition, we note that EW^4 always outperforms its preliminary versions W^4 on the data sets either before or after removing the requirement-irrelevant testing tweets. Three reasons contribute to the improvement: (1) the number of personal regions is user-specific in EW^4, which help better understand user mobility; (2) the weights of topics and regions for selecting words and locations in EW^4 are also user-specific; and (3) EW^4 is a non-parametric Bayesian model, which is more robust to overfitting.

7.4. Sample Mobility Pattern

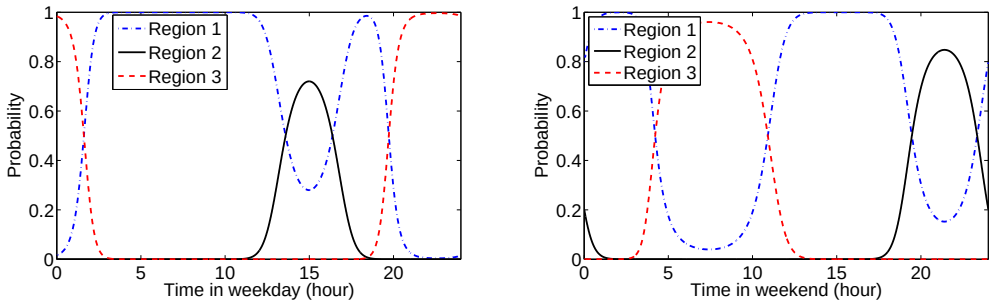
We take the model trained on the *WW* data set as an example to demonstrate the mobility pattern discovered by EW^4.

We randomly select a user, and plot her personal regions in Figure 8, and the time patterns of each region in Figure 9. Figure 8 shows that the user has three personal regions centering at different locations in the city. In addition, the contour lines of the region 3 are more concentrated than that of region 1, showing that the user usually stays in a small region at the center of region 3, but visits a relatively larger range of places around region 1.

From Figure 9, we observe that the user has different time patterns over the personal regions on weekday and weekend: *e.g.,* the user is more likely to stay at region 2 on weekday afternoon, but to stay at region 2 on weekend evening. In addition, the user is more likely to spend more time in region 1 in the daytime of weekends, but only visit region 1 at dinner time of weekdays.

7.5. Results of Example Applications

In addition to location prediction for tweets and requirement-aware location recommendation, we implement another two applications, namely, user prediction and user's location prediction. We now present their evaluation results in this subsection.



(a) Weekday (b) Weekend

Fig. 9. Region distribution over time

Table IV. Location prediction Acc of PMM, W^4 and EW^4

Acc	*WW*	*USA*
PMM	0.0423	0.1102
W^4	0.0776	0.2953
EW^4	**0.1423**	**0.5054**

Location prediction for user. This task aims to predict the location at which a given user is most likely to stay at a given time. For each tweet in the 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 employ prediction accuracy (Acc) as the evaluation metric, which shows the percentage of correct predictions.

We compare the performance of W^4 and EW^4 with a user mobility model PMM [Cho et al. 2011], on both data sets. Note that here we do not use the text of tweets. PMM is therefore applicable but not the other baselines for predicting locations of tweets using text as input [Li et al. 2011; Kinsella et al. 2011; Hong et al. 2012].

The results are reported in Table V. In location prediction, W^4 outperforms PMM by 83.45% and 167.97% on the two data sets, respectively. Potential reasons are two-fold. First, 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. Second, W^4 exploits both the functional and geographical information of locations, while PMM only utilizes the latter. Comparing to W^4, the proposed model EW^4 improves the accuracy by 83.38% and 71.15%, respectively. The improvement may come from: (1) the HDP model, which is more robust to overfitting problem; (2) user-specific number of personal regions, which enables us to model users' mobility more precisely; (3) user-specific weight between topics and regions for location and word generation, which can discover the different preferences between users.

User prediction. User prediction aims to predict the user who is most likely to visit a given location at a given time. For each tweet in the 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 evaluate the performance using prediction accuracy. Note that the PMM model proposed in [Cho et al. 2011] can also be used for user prediction, if we use location and time as input, and find the user who can maximize the likelihood. The experimental results are reported in Table V, which show that our method outperforms the baseline method significantly for similar reasons discussed earlier.

8. CONCLUSION

The availability of geo-tagged tweets enables us to design many appealing applications such as context-aware recommendation and search. To make accurate recommendations and retrieval, we need to study individuals' mobility behaviors from four factors, namely user, geographic information, time, and activity. To the best of our knowledge, our preliminary model W^4 is the only existing

Table V. User prediction Acc of PMM, W^4 and EW^4

Acc	*WW*	*USA*
PMM	0.4163	0.4021
W^4	0.5063	0.5863
EW^4	**0.5351**	**0.7679**

model that considers all of four aspects. In this article, we present its enhanced version, a novel HDP-based generative model EW^4, which is capable of jointly modeling the four factors, and providing a comprehensive description of user mobility behavior. The proposed model has a variety of applications in contextual search and recommendation. We evaluate the performance of EW^4 for several tasks on two real-world data sets, and the experimental results show that the proposed method EW^4 outperforms state-of-the-art baselines significantly for these applications.

ACKNOWLEDGMENTS

The authors would like to thank the guest editor in chief, the associate editor and the anonymous reviewers for their valuable comments, which help the authors improve this work.

REFERENCES

Amr Ahmed, Liangjie Hong, and Alexander J. Smola. 2013. Hierarchical geographical modeling of user locations from social media posts. In *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013*. 25–36. http://dl.acm.org/citation.cfm?id=2488392

Sandro Bauer, Anastasios Noulas, Diarmuid Ó Séaghdha, Stephen Clark, and Cecilia Mascolo. 2012. Talking Places: Modelling and Analysing Linguistic Content in Foursquare. In *2012 International Conference on Privacy, Security, Risk and Trust, PASSAT 2012, and 2012 International Confernece on Social Computing, SocialCom 2012, Amsterdam, Netherlands, September 3-5, 2012*. 348–357. DOI: http://dx.doi.org/10.1109/SocialCom-PASSAT.2012.107

Paul N. Bennett, Filip Radlinski, Ryen W. White, and Emine Yilmaz. 2011. Inferring and using location metadata to personalize web search. In *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*. 135–144. DOI: http://dx.doi.org/10.1145/2009916.2009938

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

Zhiyuan Cheng, James Caverlee, Kyumin Lee, and Daniel Z. Sui. 2011. Exploring Millions of Footprints in Location Sharing Services. In *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011*. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2783

Eunjoon Cho, Seth A. Myers, and Jure Leskovec. 2011. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, August 21-24, 2011*. 1082–1090. DOI: http://dx.doi.org/10.1145/2020408.2020579

Bin Cui, Hong Mei, and Beng Chin Ooi. 2014. Big data: the driver for innovation in databases. *National Science Review* 1, 1 (2014), 27–30.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A Latent Variable Model for Geographic Lexical Variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP 2010, 9-11 October 2010, MIT Stata Center, Massachusetts, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*. 1277–1287. http://www.aclweb.org/anthology/D10-1124

Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. 2008. Understanding individual human mobility patterns. *Nature* 453, 7196 (2008), 779–782.

Qiang Hao, Rui Cai, Changhu Wang, Rong Xiao, Jiang-Ming Yang, Yanwei Pang, and Lei Zhang.

2010. Equip tourists with knowledge mined from travelogues. In *Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010*. 401–410. DOI: http://dx.doi.org/10.1145/1772690.1772732

Thomas Hofmann. 1999. Probabilistic Latent Semantic Indexing. In *SIGIR '99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 15-19, 1999, Berkeley, CA, USA*. 50–57. DOI: http://dx.doi.org/10.1145/312624.312649

Liangjie Hong, Amr Ahmed, Siva Gurumurthy, Alexander J. Smola, and Kostas Tsioutsiouliklis. 2012. Discovering geographical topics in the twitter stream. In *Proceedings of the 21st World Wide Web Conference 2012, WWW 2012, Lyon, France, April 16-20, 2012*. 769–778. DOI: http://dx.doi.org/10.1145/2187836.2187940

Bo Hu and Martin Ester. 2013. Spatial topic modeling in online social media for location recommendation. In *Seventh ACM Conference on Recommender Systems, RecSys '13, Hong Kong, China, October 12-16, 2013*. 25–32. DOI: http://dx.doi.org/10.1145/2507157.2507174

Rosie Jones, Ahmed Hassan Awadallah, and Fernando Diaz. 2008. Geographic features in web search retrieval. In *Proceedings of the 5th ACM Workshop On Geographic Information Retrieval, GIR 2008, Napa Valley, California, USA, October 29-30, 2008*. 57–58. DOI: http://dx.doi.org/10.1145/1460007.1460023

Sheila Kinsella, Vanessa Murdock, and Neil O'Hare. 2011. "I'm eating a sandwich in Glasgow": modeling locations with tweets. In *Proceedings of the 3rd International CIKM Workshop on Search and Mining User-Generated Contents, SMUC 2011, Glasgow, United Kingdom, October 28, 2011*. 61–68. DOI: http://dx.doi.org/10.1145/2065023.2065039

Christoph Carl Kling, Jérôme Kunegis, Sergej Sizov, and Steffen Staab. 2014. Detecting non-gaussian geographical topics in tagged photo collections. In *Seventh ACM International Conference on Web Search and Data Mining, WSDM 2014, New York, NY, USA, February 24-28, 2014*. 603–612. DOI: http://dx.doi.org/10.1145/2556195.2556218

Yehuda Koren, Robert M. Bell, and Chris Volinsky. 2009. Matrix Factorization Techniques for Recommender Systems. *IEEE Computer* 42, 8 (2009), 30–37. DOI: http://dx.doi.org/10.1109/MC.2009.263

Chenliang Li, Aixin Sun, and Anwitaman Datta. 2012. Twevent: segment-based event detection from tweets. In *21st ACM International Conference on Information and Knowledge Management, CIKM'12, Maui, HI, USA, October 29 - November 02, 2012*. 155–164. DOI: http://dx.doi.org/10.1145/2396761.2396785

Wen Li, Pavel Serdyukov, Arjen P. de Vries, Carsten Eickhoff, and Martha Larson. 2011. The where in the tweet. In *Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24-28, 2011*. 2473–2476. DOI: http://dx.doi.org/10.1145/2063576.2063995

Qiaozhu Mei, Chao Liu, Hang Su, and ChengXiang Zhai. 2006. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *Proceedings of the 15th international conference on World Wide Web, WWW 2006, Edinburgh, Scotland, UK, May 23-26, 2006*. 533–542. DOI: http://dx.doi.org/10.1145/1135777.1135857

Kevin P Murphy. 2007. Conjugate Bayesian analysis of the Gaussian distribution. *def* 1, $2\sigma 2$ (2007), 16.

Carl Edward Rasmussen. 1999. The Infinite Gaussian Mixture Model. In *Advances in Neural Information Processing Systems 12, [NIPS Conference, Denver, Colorado, USA, November 29 - December 4, 1999]*. 554–560. http://nips.djvuzone.org/djvu/nips12/0554.djvu

Alan Ritter, Mausam, Oren Etzioni, and Sam Clark. 2012. Open domain event extraction from twitter. In *The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, Beijing, China, August 12-16, 2012*. 1104–1112. DOI: http://dx.doi.org/10.1145/2339530.2339704

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th International Conference*

on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010. 851–860. DOI: http://dx.doi.org/10.1145/1772690.1772777

Sergej Sizov. 2010. GeoFolk: latent spatial semantics in web 2.0 social media. In *Proceedings of the Third International Conference on Web Search and Web Data Mining, WSDM 2010, New York, NY, USA, February 4-6, 2010*. 281–290. DOI: http://dx.doi.org/10.1145/1718487.1718522

Chaoming Song, Tal Koren, Pu Wang, and Albert-László Barabási. 2010a. Modelling the scaling properties of human mobility. *Nature Physics* 6, 10 (2010), 818–823.

Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. 2010b. Limits of predictability in human mobility. *Science* 327, 5968 (2010), 1018–1021.

Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. 2006. Hierarchical Dirichlet Processes. *J. Amer. Statist. Assoc.* 101, 476 (2006), 1566–1581. DOI: http://dx.doi.org/10.1198/016214506000000302

Chong Wang, Jinggang Wang, Xing Xie, and Wei-Ying Ma. 2007. Mining geographic knowledge using location aware topic model. In *Proceedings of the 4th ACM Workshop On Geographic Information Retrieval, GIR 2007, Lisbon, Portugal, November 9, 2007*. 65–70. DOI: http://dx.doi.org/10.1145/1316948.1316967

Benjamin Wing and Jason Baldridge. 2011. Simple supervised document geolocation with geodesic grids. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*. 955–964. http://www.aclweb.org/anthology/P11-1096

Jinyun Yan, Wei Chu, and Ryen W. White. 2014. Cohort modeling for enhanced personalized search. In *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast , QLD, Australia - July 06 - 11, 2014*. 505–514. DOI: http://dx.doi.org/10.1145/2600428.2609617

Mao Ye, Peifeng Yin, Wang-Chien Lee, and Dik Lun Lee. 2011. Exploiting geographical influence for collaborative point-of-interest recommendation. In *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*. 325–334. DOI: http://dx.doi.org/10.1145/2009916.2009962

Zhijun Yin, Liangliang Cao, Jiawei Han, Chengxiang Zhai, and Thomas S. Huang. 2011. Geographical topic discovery and comparison. In *Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April 1, 2011*. 247–256. DOI: http://dx.doi.org/10.1145/1963405.1963443

Quan Yuan, Gao Cong, Zongyang Ma, Aixin Sun, and Nadia Magnenat-Thalmann. 2013a. Time-aware point-of-interest recommendation. In *The 36th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '13, Dublin, Ireland - July 28 - August 01, 2013*. 363–372. DOI: http://dx.doi.org/10.1145/2484028.2484030

Quan Yuan, Gao Cong, Zongyang Ma, Aixin Sun, and Nadia Magnenat-Thalmann. 2013b. Who, where, when and what: discover spatio-temporal topics for twitter users. In *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, Chicago, IL, USA, August 11-14, 2013*. 605–613. DOI: http://dx.doi.org/10.1145/2487575.2487576

Quan Yuan, Gao Cong, and Aixin Sun. 2014. Graph-based Point-of-interest Recommendation with Geographical and Temporal Influences. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, November 3-7, 2014*. 659–668. DOI: http://dx.doi.org/10.1145/2661829.2661983