Spatial Topic Modeling in Online Social Media for Location Recommendation

Bo Hu
School of Computing Science
Simon Fraser University
Burnaby, BC, Canada
boh@cs.sfu.ca

Martin Ester
School of Computing Science
Simon Fraser University
Burnaby, BC, Canada
ester@cs.sfu.ca

ABSTRACT

Mobile networks enable users to post on social media services (e.g., Twitter) from anywhere. The activities of mobile users involve three major entities: user, post, and location. The interaction of these entities is the key to answer questions such as **w**ho will post a message **w**here and on **w**hat topic? In this paper, we address the problem of profiling mobile users by modeling their activities, i.e., we explore topic modeling considering the spatial and textual aspects of user posts, and predict future user locations. We propose the first **ST** (**S**patial **T**opic) model to capture the correlation between users' movements and between user interests and the function of locations. We employ the sparse coding technique which greatly speeds up the learning process. We perform experiments on two real life data sets from Twitter and Yelp. Through comprehensive experiments, we demonstrate that our proposed model consistently improves the average precision@1,5,10,15,20 for location recommendation by at least 50% (Twitter) and 300% (Yelp) against existing state-of-the-art recommendation algorithms and geographical topic models.

Categories and Subject Descriptors

H.2 [**Database Management**]: Database Applications-Data Mining

General Terms

Algorithms, Design, Measurement, Experimentation

Keywords

Spatial Topic Model, Mobile Users, Location Recommendation

1. INTRODUCTION

With the rapid growth of mobile network users, the way users consume Web 2.0 is changing substantially. Mobile networks enable users to post on social media services (e.g., Twitter or Yelp) from anywhere. This new phenomenon led to the emergence of a new line of research to mine the behavior of social media users taking into account the spatial aspects of their engagement with online social media.

The activities of mobile users can typically be represented as follows: a user appears at a certain location (with a pair of latitude and longitude coordinates), and leaves a post (e.g., tweet or review) which is likely semantically related to the user and/or the location. These activities involve three major entities: user, post, and location. The interaction of these entities is the key to answer questions such as **w**ho will post a message **w**here and on **w**hat topic? In this paper, we address the problem of profiling mobile users by modeling their activities, i.e., we explore topic modeling considering the spatial and textual aspects of user posts, and predict future user locations.

Several works in the literature have addressed some of the above aspects. In recommender systems, [20, 2, 13] have proposed probabilistic matrix factorization models mining latent user and location preferences to predict user locations, but they totally ignore one of the key components: user posts. Another line of works [18, 21] has focused on user posts and proposed topic models to analyze geographical topics. Most recently, Hong et. al. [10] proposed a geographical topic model to capture language patterns of different regions and different users. Note that the users' distributions over regions are assumed to be independent from each other.

We observe that user movements sometimes correlate if two users have similar lifestyle or living routine. For example, many students from New York University live in the same neighborhood near the campus, and their movement trajectories correlate to each other. They may go to the same restaurants, coffee shops and grocery stores. Therefore, we argue that considering the movements of different users independently as in [10] is not the best way, and that we can predict a user's movement more accurately taking into account the movements of similar users. This idea underlies the paradigm of collaborative filtering.

A second observation is that user interest affects user movement not at the "syntactic" level of 2-dimensional coordinates but at the "semantic" level of places with a certain function. Existing spatial topic models with 2-dimensional coordinates do not distinguish the following two scenarios: 1) two users appear in the same location, like a hockey themed bar, and 2) two users appear in two different locations that are adjacent to each other, where one is a hockey

RecSys'13, October 12–16, 2013, Hong Kong, China.

http://dx.doi.org/10.1145/2507157.2507174.

themed bar and the other one is a facial salon. Intuitively, male users who are interested in sports often go to sport bars and watch games, while female users often go to facial salons. Two users in the first scenario share the same interest, while two users have totally different interests in the second scenario. As a result, without considering the fact that user movements are influenced not only by the coordinates of a location but also by its function, the predictive ability of the model will be greatly reduced.

Motivated by the above observations, this paper explores the following two questions:

1. How are user movements correlated to each other?

2. How does user interest affect user movement at the "semantic" level of locations?

We propose a spatial topic model, called **ST** (**S**patial **T**opic), that takes the correlation of users' movements, and the correlation of user movement and user interest into account. As in existing models, a post is represented as an unordered collection of words (a bag-of-words assumption) associated with user and location, which are all considered as observed random variables. Different from existing works [18, 21, 10], a location in this paper is defined as a place with a semantic functionality and with its 2D coordinates. A set of latent random variables is also defined, i.e., regions and topics are latent, and each post is assigned to a region and a topic. We assume that each location is assigned to one and only one region, and its coordinates are generated by a 2-dimensional Gaussian distribution. For example, in New York City, regions could be areas that corresponded to community districts, such as Manhattan, Brooklyn, and Queens etc. Different from existing models, in order to generate a location of a post by a particular user, the model considers the user's interest and the locations of "similar" users. We develop a MCEM (Monte Carlo Expectation Maximization) method to learn the latent random variables and parameters that maximize the likelihood of the observed random variables, and the sparse coding technique is used to improve the efficiency of the learning method.

We perform experiments on two real life data sets from Twitter and Yelp. All posts (tweets and reviews) in the data sets are annotated with corresponding users and locations. We evaluate the effectiveness of our proposed model and of state-of-the-art models in terms of accuracy of location prediction, i.e., given a post and its author, we recommend top-k locations to the user.

The major contributions of this paper are as follows:

- We propose the first spatial topic model to capture the correlation between users' movements and between user interests and the function of locations.

- We employ the sparse coding technique which greatly speeds up the learning process.

- Through comprehensive experiments, we demonstrate that our proposed model consistently improves the average precision@1,5,10,15,20 for location recommendation by at least 50% (Twitter) and 300% (Yelp) compared to existing state-of-the-art recommendation algorithms and geographical topic models.

2. RELATED WORK

In this section, we briefly review related work. There are three lines of related work, which are geographical topic modeling, location recommendation, and user movement analysis.

Geographical Topic Modeling. Topic modeling is a classic task to enable text analysis at a semantic level. A topic model assumes that each document in a given data set is associated with a topic distribution, and each topic with a word distribution. The most representative models are PLSA [9] and LDA [1]. Recently, there are many works [19, 18, 21, 10] in the area of geographical topic modeling, which detect geographical regions and topics from documents that are associated with locations.

Yin et al. [21] and Eisenstein et al. [7] propose similar models, where the coordinates in each document are drawn from a 2D Gaussian distribution and the region is drawn from a Multinomial distribution over all regions.

Recently, Hong et al. [10] propose a model, called GT (Geographical Topic), assuming that each user has a distribution over all regions, i.e., users tend to appear in a small subset of all regions, and each user has a distribution over all topics, i.e., users tend to have different interests on different topics. They conduct extensive experiments on a large scale Twitter data set and their model achieves better location prediction performance than existing models.

Although all the above works discover regions and geographical topics, they do not consider the correlation of users' movements. Additionally, they assume that users' interests influence users' regional preferences directly but influence users' locations only indirectly. Finally, these models ignore the functionality of locations, which greatly reduces the predictive ability of the model as will be shown in our experiments.

Location Recommendation. In the classic framework of recommender systems, we have a user-item matrix and each element in the matrix represents the user's rating of that item. To put it in the context of location recommendation, locations can be viewed as items, and binary user item ratings can represent whether a user has visited the corresponding locations. Particularly, the user-location matrix can be computed by the user and location latent factors through MF (Matrix Factorization) techniques [11, 12]. Recent works [20, 22, 2, 13] extend PMF (Probabilistic Matrix Factorization) [16] for location recommendation by considering the distance between users and locations, i.e., closer locations are recommended with higher probabilities.

The authors of [2] have proposed a MF model considering geographical influence for POI (Point-Of-Interest) recommendation in location-based social networks. Their model detects multiple centers for each user based on their history of locations, and each center has 2-dimensional coordinates. Motivated by the effect of geographical influence, the probability of recommending a location is inversely proportional to the distance between the location and the nearest user center. Similarly, a recent work [13] has proposed GLDA (Geo Latent Dirichlet Allocation) to capture users' location preferences by combining LDA and geographical influence.

Note that all these methods ignore the text of posts which is an essential component in our problem definition.

User Movement Analysis. Some works [8, 3, 4, 5] have been studying user movements in location-based social networks. Since most user posts are not associated with coor-

dinates, these works address the following problem: given a set of geo-tagged posts from many users, learn a model of region specific words, and apply this model to predict the user location of un-tagged posts based on their content. This is different from the problem in this paper, which is to predict the location of posts by users with many geographical posts. Cheng et al. [3, 4] develop probabilistic methods to identify local words in tweets, and they predict user locations based on the local words in their tweets. Similarly, [8] proposes a Multinomial Naive Bayes model to predict the Twitter user profile's location at the granularity of the city level.

[5] studies the problem of modeling human mobility in social networks, and one of their interesting findings is that users tend to move within a small number of regions, e.g., around their home and office. Another interesting finding is that a user's movement trajectory correlates to that of their friends. Furthermore, a recent work [15] presents a probabilistic model incorporating social networks and achieves better performance for tweet location prediction. This type of work focuses on the impact of social networks on user movements, while we do not use social networks.

3. SPATIAL TOPIC MODEL

In this section, we first introduce the problem definition and then present our proposed **ST** (**S**patial **T**opic) model.

3.1 Problem Definition

On social media sites, such as Twitter or Yelp, a large number of users generate content. These user generated posts can consist of personal information, news, comments or reviews. We are in particular interested in the text and location of the posts. More precisely, we assume that a post has the following attributes: text, author, and location. An example of publicly available tweets is as follows:

- Close to the equator, perfect soil and high elevation make @CafeDAltamira produce the perfect cup of Coffee!! # Honduras # Coffee | @XXX | 37.38 -121.90 | San Jose | CA | United States | 0befbacea94beb06.

This tweet states that a Twitter user from San Jose, California compliments the coffee at a coffee shop, where "37.38, -121.90" are the latitude and longitude coordinates, and "0befbacea94beb06" is the unique label of the coffee shop.

We assume that all the documents are authored by a user from a fixed set of size U and all the words are from a fixed vocabulary of size V. We associate each user with a set of posts, and the number of posts of user u is denoted as D_u. Each post is represented by a set of words (the number of its words is denoted as $N_{u,d}$), and a pair of latitude and longitude coordinates. For convenience, we consider "tweet", "review", "post" and "document" as synonyms in this paper. Formally, a document d is defined by $d = \{\mathbf{w}, u, i\}$, where w, u, i represents set of (index of) words, the index of user and location respectively. l_i represents the coordinates of location i. A document collection $\mathcal{D}$ is defined as a set of documents from all users. We assume that there is a set of latent topics and a set of latent regions in the document collection $\mathcal{D}$. Each document d is assigned to one of the topics z_d and regions r_d. We use Z and R to denote the number of topics and number of regions, respectively.

A semantically coherent topic in the document collection $\mathcal{D}$ is associated with a probability distribution over all words in the vocabulary, and a probability distribution over all locations. A region has a geographical center, and it is comprised of a set of documents, which are coherent in topics and close to the center geographically. We assume that different users show different distributions over topics and regions. All notations described above are listed in Table 1.

Table 1: Notations of input and output data

Variable	Interpretation
$w_{u,d,n}$	n^{th} word of the d^{th} document posted by the u^{th} user
$i_{u,d}$	location index of the d^{th} document posted by the u^{th} user
l_i	latitude and longitude coordinates of the i^{th} location
$z_{u,d}$	topic assignment of the d^{th} document posted by the u^{th} user
$r_{u,d}$	region assignment of the d^{th} document posted by the u^{th} user
Z	number of topics
R	number of regions
U	number of users
I	number of locations
D_u	number of documents of user u
$N_{u,d}$	number of words in document d of user u
V	size of the vocabulary

Based on the above definitions, we formalize our research problem as follows:

PROBLEM 1 (SPATIAL TOPIC MODELING). *Given a document collection $\mathcal{D}$, and numbers Z of topics and R of regions, the task is to model and extract a set of topics and a set of regions.*

3.2 Model

To address our research problem, we introduce the ST model. Figure 1 shows the graphical model of ST.

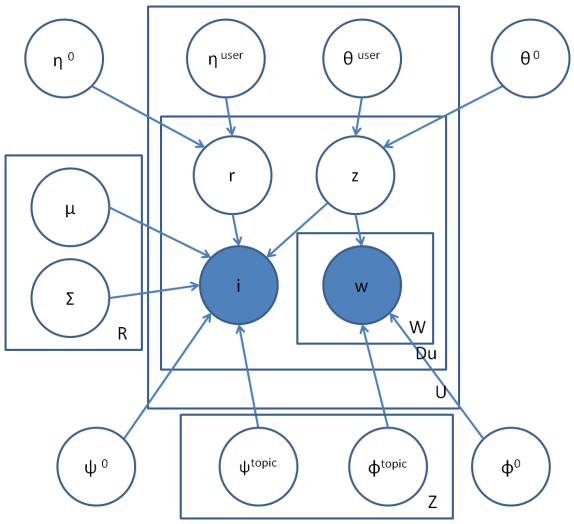


Figure 1: The graphical model of ST

We first introduce the notations of our model and listed in Table 1. Our input data, i.e., words and locations, are modeled as observed random variables, shown as shaded circles in Figure 1, and we use $w_{u,d,n}$ and $i_{u,d}$ to denote them.

l_i is a pair of latitude and longitude real values of i^{th} location. Similar to existing models as in [21, 10], the topic and region index of documents are considered as latent random variables, which are denoted as $z_{u,d}$ and $r_{u,d}$ respectively. Users are associated with topic and region distributions, i.e., θ^{user} and η^{user}, from which the topics and regions of posts are sampled. Topics are associated with word distributions ϕ^{topic}. Given the sampled topic, words are drawn from the word distribution of that topic. The background distributions of words, topics, and regions are denoted as ϕ^0, θ^0, and η^0. All parameters are listed in Table 2.

Table 2: Notations of parameters

Variable	Interpretation
θ^0	topic distribution of the background
θ_u^{user}	topic distribution of the u^{th} user
ϕ^0	word distribution of the background
ϕ_z^{topic}	word distribution of the z^{th} topic
η^0	region distribution of the background
η_u^{user}	region distribution of the u^{th} user
ψ^0	location distribution of the background
ψ_z^{topic}	location distribution of the z^{th} topic
μ_r	region mean location of the r^{th} region
Σ_r	region location covariance of the r^{th} region

An important change from existing models is that instead of generating the coordinates of posts, ST generates the index of the location of posts. Another major change is that, to model the impact of user interest on user movement, ST assumes that the location depends not only on the region but also on the topic. Consequently, it adds a location distribution ψ^{topic} for each topic. Existing models assume that the 2D Gaussian distribution with center μ and covariance Σ of the sampled region governs the choice of locations visited, i.e., the closer a location to the center, the higher the probability of visiting that location, and ST has the same assumption. Additionally, ST assumes that another important reason why the user visits the location can be attributed to the user interests. Since locations with different functions can have very similar coordinates, this assumption is much more meaningful when considering "semantic" locations.

Particularly, users have different topic distributions, and topics have different location distributions, so that the dependency between user interests and user locations is transferred through the topic. ST captures the correlation between movements of different users, such that users who have similar movements share the same topics. Note that the topics serve a similar role as the latent factors in MF (Matrix Factorization). Different from the existing MF methods [2, 13], ST associates a word distribution with a topic so that it can describe the latent user and location factors. Intuitively, collaborative filtering assumes that locations A and B should both have high probabilities in location distributions of some topic(s) in our case if many users frequently co-occur in both A and B. Location A and B do not necessarily have the same functionality. However, ST further assumes that locations with high probabilities for the same topic should be cohesive in their functions, e.g., a topic with high probabilities for words like "coffee", "Java", and "mocha" should have high probability only for coffee shops. This design enables ST to detect users with similar interests and locations with similar functions, and enables ST to better deal with "cold start" users, i.e., users who have very few posts, since the words of their few posts are more informative than their locations.

Next, we describe the generative process of the ST model for a single document d.

- Draw a region index $r_{u,d}$
  - $r_{u,d} \sim p(r_{u,d}|u, \eta^0, \eta^{user})$
- Draw a topic index $z_{u,d}$
  - $z_{u,d} \sim p(z_{u,d}|u, \theta^0, \theta^{user})$
- Draw a location index $i_{u,d}$, given the region index $r_{u,d}$ and topic index $z_{u,d}$
  - $i_{u,d} \sim p(i_{u,d}|r_{u,d}, z_{u,d}, \psi^0, \psi^{topic}, \mu, \Sigma)$
- Draw each word in d given the topic index $z_{u,d}$
  - $w_{u,d,n} \sim p(w_{u,d,n}|z_{u,d}, \phi^0, \phi^{topic})$

For each document, the ST model generates the location and words consecutively. To generate a location, the model first samples a region from the set of regions. To generate a region r, we use a multinomial distribution as follows:

$$p(r_{u,d}|u, \eta^0, \eta^{user}) = p(r_{u,d}|\eta^0 + \eta_u^{user}) \tag{1}$$

where η^0 is the global distribution of regions and η_u^{user} is the region distribution of user u. To simplify the notations, we use $p(r|\eta^0+\eta_u^{user}) = \beta_{u,r}$ as shown in Equation 6 (see Figure 2). This approach employs the sparse coding technique introduced in the SAGE (Sparse Additive Generative) model [6]. The major advantage of SAGE is that it does not require additional latent "switching" variables when the model needs to take multiple factors into account. For example, in order to model topics, based on the background word distribution, for each topic SAGE models the difference in log-frequencies from the background word distribution instead of the log-frequencies themselves.

Each location i is drawn depending on its corresponding region r and corresponding topic z. Given the sampled topic z and sampled region r, ST draws the location $i_{u,d}$ as follows:

$$\begin{aligned} i_{u,d} \sim \quad & p(i_{u,d}|r_{u,d}, z_{u,d}, \psi^0, \psi^{topic}, \mu, \Sigma) \\ & = p(i_{u,d}|\psi^0 + \psi_{z_{u,d}}^{topic}) \times p(l_{i_{u,d}}|\mu_{r_{u,d}}, \Sigma_{r_{u,d}}) \end{aligned} \tag{2}$$

where $p(i|\psi^0 + \psi_z^{topic}) = \delta_{z,i}$ is shown in Equation 6, and $p(l_i|\mu_r, \Sigma_r) = \mathcal{N}(l_i|r, \mu, \Sigma)$, which is the PDF of the Multivariate Gaussian distribution. This is the product of the probability of drawing the coordinates of the location from the 2D Gaussian distribution μ_r, Σ_r of that region, and the probability of drawing the index of the location from the location distribution ψ_z^{topic} of that topic.

Similarly, for generating the topic and word index, the model uses a multinomial distribution considering the background and user topic distributions together, and the background and topic word distributions together, respectively as follows:

$$p(z_{u,d}|u, \theta^0, \theta_u^{user}) = p(z_{u,d}|\theta^0 + \theta_u^{user}) \tag{3}$$

$$p(w_{u,d,n}|z_{u,d}, \phi^0, \phi^{topic}) = p(w_{u,d,n}|\phi^0 + \phi_{z_{u,d}}^{topic}) \tag{4}$$

where $p(z|\theta^0+\theta_u^{user}) = \alpha_{u,z}$ and $p(w_{u,d,n}|\phi^0+\phi_z^{topic}) = \gamma_{z,w}$ are shown in Equation 6.

$$p(\mathbf{z},\mathbf{r},\mathbf{w},\mathbf{i}|\Theta) = p(\mathbf{z}|\theta^0,\theta^{user}) \times p(\mathbf{r}|\eta^0,\eta^{user}) \times p(\mathbf{w}|\mathbf{z},\phi^0,\phi^{topic}) \times p(\mathbf{i}|\mathbf{r},\mathbf{z},\mu,\Sigma,\psi^0,\psi^{topic})$$
$$= \prod_{u=1}^{U}\prod_{d=1}^{D_u}\alpha_{u,z_{u,d}} \times \prod_{u=1}^{U}\prod_{d=1}^{D_u}\beta_{u,r_{u,d}} \times \prod_{u=1}^{U}\prod_{d=1}^{D_u}\prod_{n=1}^{N_{u,d}}\gamma_{z_{u,d},w_{u,d,n}} \times \prod_{u=1}^{U}\prod_{d=1}^{D_u}\delta_{z_{u,d},i_{u,d}} \times \prod_{u=1}^{U}\prod_{d=1}^{D_u}\mathcal{N}(l_{i_{u,d}}|r_{u,d},\mu,\Sigma) \quad (5)$$

$$\alpha_{u,z} = \frac{\exp(\theta^0_z + \theta^{user}_{u,z})}{\sum_{zz=1}^{Z}\exp(\theta^0_{zz} + \theta^{user}_{u,zz})}, \beta_{u,r} = \frac{\exp(\eta^0_r + \eta^{user}_{u,r})}{\sum_{rr=1}^{R}\exp(\eta^0_{rr} + \eta^{user}_{u,rr})}, \gamma_{z,w} = \frac{\exp(\phi^0_w + \phi^{topic}_{z,w})}{\sum_{ww=1}^{V}\exp(\phi^0_{ww} + \phi^{topic}_{z,ww})}, \delta_{z,i} = \frac{\exp(\psi^0_i + \psi^{topic}_{z,i})}{\sum_{ii=1}^{I}\exp(\psi^0_{ii} + \psi^{topic}_{z,ii})} \quad (6)$$

Figure 2: The joint probability of random variables given parameters in the ST model

3.3 Parameter Learning

Our goal is to learn parameters that maximize the marginal log-likelihood of the observed random variables $\mathbf{i}, \mathbf{w}$. The marginalization is performed with respect to the latent random variables $\mathbf{z}, \mathbf{r}$, and it is hard to be maximized directly. Therefore, we apply the MCEM (Monte Carlo Expectation Maximization) algorithm to maximize the complete data likelihood $p(\mathbf{z},\mathbf{r},\mathbf{w},\mathbf{i}|\Theta)$ in Equation 5 (see Figure 2), where $\Theta = \{\mu, \Sigma, \theta^0, \theta^{user}, \phi^0, \phi^{topic}, \eta^0, \eta^{user}, \psi^0, \psi^{topic}\}$.

According to the MCEM method, we sample the latent variables $\mathbf{r}, \mathbf{z}$ in the E step and maximize the parameters Θ in the M step. To sample a single variable $r_{u,d}$ given all other variables fixed, we use Equation 7. After $\mathbf{r}$ is sampled, we sample $z_{u,d}$ similarly according to Equation 8.

$$p(r_{u,d}|\mathbf{z},\mathbf{r}_{-u,d},\mathbf{w},\mathbf{i},\Theta) \propto \beta_{u,r_{u,d}} \times \delta_{z_{u,d},i_{u,d}} \times \mathcal{N}(l_i|r,\mu,\Sigma) \quad (7)$$

$$p(z_{u,d}|\mathbf{z}_{-u,d},\mathbf{r},\mathbf{w},\mathbf{i},\Theta) \propto \alpha_{u,z_{u,d}} \times \prod_{n=1}^{N_{u,d}}\gamma_{z_{u,d},w_{u,d,n}} \times \delta_{z_{u,d},i_{u,d}} \times \mathcal{N}(l_i|r,\mu,\Sigma) \quad (8)$$

Figure 3: The sampling formulas for latent variables r,z in the ST model

In the M step, fixing all the latent variables $\mathbf{r}, \mathbf{z}$ that are sampled in the E step, we maximize the log likelihood of Equation 5 with respect to the parameters Θ. For variables μ and Σ, to obtain the maximum likelihood estimate, we take the derivative of its log likelihood with respect to μ_r and Σ_r, and set it to zero. Only one term in Equation 5 contains μ_r, so we use Equation 9 to update μ_r, where $\mathcal{I}(.)$ is an identity function, i.e. one where $r_{u,d}$ equals to r and zero otherwise, and $d(r)$ represents the number of documents assigned to region r. μ_r denotes the mean coordinates of locations of the documents assigned to region r in the E step. We use Equation 10 to update the parameter Σ_r.

$$\mu_r = \frac{1}{d(r)}\sum_{u=1}^{U}\sum_{d=1}^{D_u}\mathcal{I}(r_{u,d} == r) l_{i_{u,d}} \quad (9)$$

$$\Sigma_r = \frac{1}{d(r)-1}\sum_{u=1}^{U}\sum_{d=1}^{D_u}\mathcal{I}(r_{u,d} == r)(l_{i_{u,d}} - \mu_r)^T(l_{i_{u,d}} - \mu_r) \quad (10)$$

To update the other parameters, we use the gradient descent learning algorithm PSSG (Projected Scaled Sub-Gradient) [17], which is designed to solve optimization problems with L1 regularization on the parameters. More importantly, PSSG is scalable because it uses the quasi-Newton strategy with line search that is robust to common functions. According to the limited-memory BFGS [14] updates for the quasi-Newton method, the partial derivative functions of the parameters η^0, η^{user} are provided in the following Equations 11 and 12, where $d(u,r)$ represents the number of documents assigned to region r by user u, and $d(u)$ represents the number of documents by user u.

$$\frac{\partial L}{\partial \eta^0_r} = \sum_{u=1}^{U} d(u,r) - \sum_{u=1}^{U}\Big(d(u) \times \beta_{u,r}\Big) \quad (11)$$

$$\frac{\partial L}{\partial \eta^{user}_{u,r}} = d(u,r) - d(u) \times \beta_{u,r} \quad (12)$$

Similarly, we get derivative functions for the remaining parameters, which are omitted because of the page limit.

3.4 Location Recommendation

The ST model can be employed for location recommendation as follows. Given a document with a user, our task is to recommend top-k "new " locations, i.e., the locations that the user has not visited in the training data set, which that user will visit. More precisely, given the words and author of a document d, the probability that author u visits location i is computed as in Equation 13:

$$p(i|\mathbf{w},\Theta) \propto \sum_{r}^{R}\sum_{z}^{Z} p(\mathbf{w},i,z,r|\Theta)$$
$$= \sum_{r}^{R}\sum_{z}^{Z} p(z|\theta^0,\theta^{user}) \times p(r|\eta^0,\eta^{user}) \times p(\mathbf{w}|z,\phi^0,\phi^{topic}) \times p(i|z,r,\mu,\Sigma,\psi^0,\psi^{topic}) \quad (13)$$

We rank the locations in descending order of $p(i|\mathbf{w},\Theta)$.

4. EXPERIMENTS

In this section, we experimentally evaluate the effectiveness of the ST (Spatial Topic) model, and we compare it against some baseline methods, one of the state-of-the-art location recommendation methods [13], and one of the state-of-the-art geographical topic models [10]. We report our experimental results on Twitter and Yelp data sets, using the top-k average precision of location recommendation for measuring the quality.

4.1 Data Sets

We report our experimental results on a Twitter data set downloaded from [4][1]. We extract a data set from a rep-

[1]http://infolab.tamu.edu/data/

resentative city in the US: NYC (New York City), where all tweets contain a location label and geographical coordinates. To determine the coordinates of the location, we use the mean of the coordinates of all tweets associated with a location. Hence each location corresponds to a unique mean coordinate, and each tweet of that location has the same coordinates. Another data set is from Yelp, and it is publicly available[2]. It is from a US city – Phoenix. In the Yelp data set, each review has a location (being reviewed) that is associated with a unique pair of latitude and longitude coordinates. Note that Twitter users often check in at the same location multiple times, while Yelp users write reviews for a location only once.

In the pre-processing steps, texts are processed by tokenizing on whitespace and punctuations, while we remove the URLs starting with "http" and user names starting with "@". Then we remove all texts with non-latin characters, followed by removing stop words, and the words with occurrences less than 100. To reduce noise, we remove both users and locations with less than 10 posts. Some statistics about the data sets are presented in Table 3.

Table 3: Statistics of data sets from New York City on Twitter and Phoenix on Yelp.

#	Twitter	Yelp
Unique users	9,508	3,963
Posts	607,885	107,981
Locations	3,518	2,951
Avg. posts/user	64.93	27.24
Avg. posts/location	172.79	36.59

Note that our data sets are much larger than the ones used in [13]. Another related work [10] uses data sets from all over the world, while our data sets are at the city level. From this point of view, the size of these two data sets is comparable or larger than the ones in [10].

4.2 Experimental Setup

In our data sets, we randomly select 70% of observed data for each user as the training data, and the remaining 30% as the test data. We focus on the task of location recommendation for users based on each document, which is by far the most commonly used performance measure for spatial topic model in the literature [20, 2, 13]. In particular, we train models in the training data set, and recommend the locations based on posts by users in the test data set.

Evaluation Metric. Precision@k (top-k average precision) is used to evaluate the methods as follows. The top-k precision for a test post is $\frac{1}{k}$ if its location is among the top-k recommendations, and zero otherwise. The precision@k is the average top-k precision over all test posts.

Comparison Partners. In our experiments, we evaluate the following comparison partners, which all model (can predict) either the coordinates or index of locations:

- PMF (Probabilistic Matrix Factorization). This is a well-known model in matrix factorization in [16].
- $GLDA$ (Geo Latent Dirichlet Allocation). This is the modified LDA model, which is one of the state-of-the-art methods for location recommendation proposed in [13].
- GT (Geographical Topic). This is one of the state-of-the-art geographical topic models proposed in [10].
- $ST_{location}$ (ST_{loc} for short). This is a simplified version of the ST model, where we remove the posts, and the only observed variable is the index of locations **i** and the only latent variable is the topic **z**. Note that this model is equivalent to an LDA model that generates index of locations instead of words.
- $ST_{coordinate}$ (ST_{coo} for short). This is a simplified version of the ST model. Similar to ST_{loc}, we remove the posts from the data. Instead of generating the index of locations, ST_{coo} generates the coordinates of locations **l**, and the only latent variable is the region **r**.
- $ST_{coordinate+location}$ ($ST_{loc+coo}$ for short). This is another simplified version of the ST model, that generates both the coordinates and index of locations, and the latent variables are the topic **z** and region **r**. The only difference between this model and the full ST model is the lack of words.
- ST. This is the spatial topic model proposed in this paper.

Note that there are other existing models [19, 18, 21] proposed for geographical topic modeling. We do not compare against them because the GT model proposed in [10] is a generalization of the existing models, and it performs better than the existing models in terms of location prediction in the experiments of [10]. We do not compare against [2], since it is similar to GLDA, which is the most recent work [13] on location recommendation.

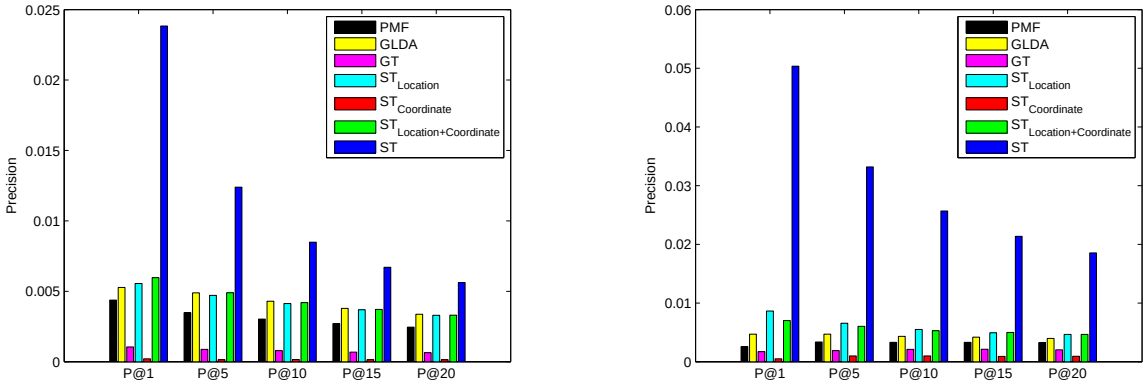
4.3 Experimental Results

For location recommendation, Figure 4(a) and 4(b) show the precision@1,5,10,15,20 results of the comparison partners in the Twitter and Yelp data sets. Note that the number of topics and regions is set to 30 and 20. We observe that our ST model consistently and drastically outperforms all other models on both data sets. Compared to the state-of-the-art methods, GLDA and GT, in the areas of recommender systems and geographical topic modeling, ST improves the precision@20 by 50% (Twitter) and 300% (Yelp), and the gain is even higher for smaller values of k. This indicates that modeling the user interests and the correlation of user movements can help improve the accuracy of location recommendation.

We also observe that the precision difference between ST and other models on Yelp is much larger than on Twitter. We argue that this is because 1) the posts on Yelp are much longer than on Twitter; 2) the words used on Yelp are more formal than on Twitter. As a result, it is easier to capture the user interests on Yelp than on Twitter.

We further analyze the contributions of different components in ST, by comparing the performance of ST and its simplified versions: ST_{loc}, ST_{coo}, and $ST_{loc+coo}$. We observe that modeling the index (semantics) of locations in ST_{loc} is much more precise than modeling the coordinates of locations in ST_{coo}. Comparing ST and $ST_{loc+coo}$, we see that the user interests expressed in the posts indeed enable more accurate location recommendation. Furthermore, we observe that $ST_{loc+coo}$ clearly outperforms ST_{coo}, demonstrating the contribution of exploiting the correlation of user movements.

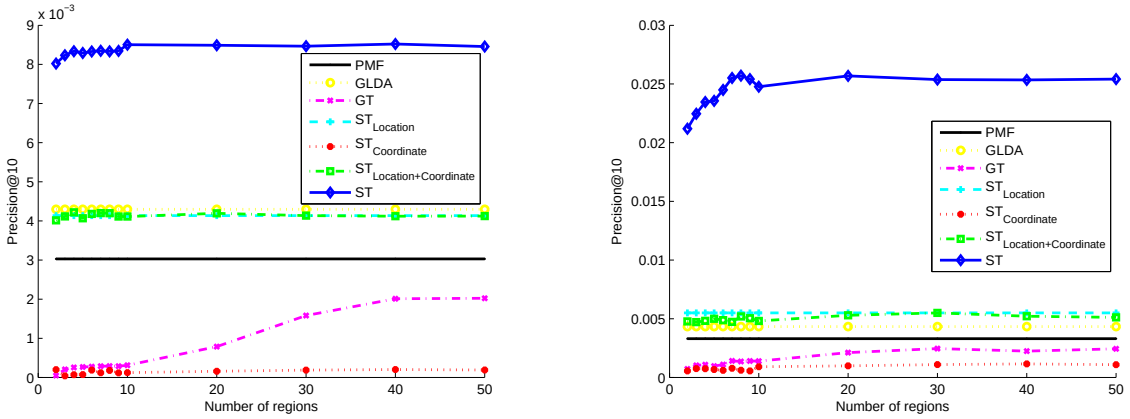
[2]https://www.yelp.com/dataset_challenge/



(a) Twitter data set.

(b) Yelp data set.

Figure 4: Precision@1,5,10,15,20 of comparison partners.



(a) Twitter data set.

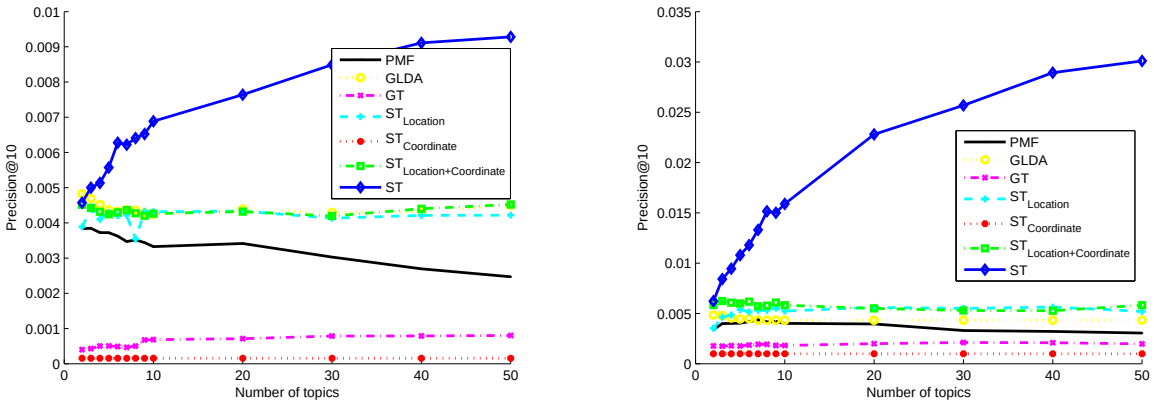
(b) Yelp data set.

Figure 5: Precision@10 of the comparison partners for different number of regions. The number of topics is set to 30.

To analyze the impact of the input parameters, we show the precision@10 of the comparison partners for different numbers of regions (see Figure 5(a) and 5(b)) and topics (see Figure 6(a) and 6(b)). The results for precision@1,5,15,20 are similar to the results for precision@10. We observe that ST consistently outperforms the other comparison partners for all number of regions and topics. Furthermore, as the number of regions increases, the precision@10 of ST and GT increases and reaches a peak at first, and it plateaus when the number of regions reaches 10 or 20. Similarly, as the number of topics increases, the precision of ST increases. Some models, such as PMF, $GLDA$ and ST_{loc}, do not take the number of regions as their input, so that their precision is constant in Figure 5. Overall, the results of ST are relatively robust to the choice of the input parameters.

5. CONCLUSION AND FUTURE WORK

In this paper, we address the problem of spatial topic modeling in online social media, such as Twitter and Yelp, for user-generated content with location. Previous work has explored topic models and recommendation algorithms that model either user and location, or user and post, but they do not consider all of them together. We propose the first spatial model to capture spatial and textual aspects of posts, as well as user profiles in a single topic model, called Spatial Topic (ST) model. ST exploits the interdependencies between user movements, and between user interests and user movements. More specifically, ST is based on the intuition that 1) users' movements correlate with each other; 2) users' interests affect the movements of users. We argue that taking the correlation of users' movements, and the correlation of user movement and user interest into account enables a more accurate discovery of relevant regions and topics. We present the graphical model of ST and a corresponding method of parameter learning. We perform an experimental evaluation on Twitter and Yelp data sets from New York City and Phoenix. We compare ST against a state-of-the-art geographical topic model and a state-of-the-art recommendation method in terms of location recommendation. Our



(a) Twitter data set.

(b) Yelp data set.

Figure 6: Precision@10 of the comparison partners for different number of topics. The number of regions is set to 20.

experiments demonstrate drastically improved performance in location recommendation.

An interesting direction for future work is to integrate other aspects that may impact user locations, i.e. ratings and time, into the spatial topic model. Open questions are in particular how ratings of locations attract users, as well as if and how time (hour, day, week, year) influences user locations.

6. REFERENCES

[1] D. M. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

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

[3] Z. Cheng, J. Caverlee, and K. Lee. You are where you tweet: a content-based approach to geo-locating twitter users. CIKM, pages 759–768, 2010.

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

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

[6] J. Eisenstein, A. Ahmed, and E. P. Xing. Sparse additive generative models of text. ICML, pages 1041–1048, 2011.

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

[8] B. Hecht, L. Hong, B. Suh, and E. H. Chi. Tweets from justin bieber's heart: the dynamics of the location field in user profiles. In *CHI*, pages 237–246, 2011.

[9] T. Hofmann. Probilistic latent semantic analysis. In *UAI*, 1999.

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

[11] Y. Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. KDD, pages 426–434, 2008.

[12] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42:30–37, 2009.

[13] T. Kurashima, T. Iwata, T. Hoshide, N. Takaya, and K. Fujimura. Geo topic model: joint modeling of user's activity area and interests for location recommendation. WSDM, pages 375–384, 2013.

[14] D. C. Liu, J. Nocedal, D. C. Liu, and J. Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical Programming*, pages 503–528, 1989.

[15] A. Sadilek, H. Kautz, and J. P. Bigham. Finding your friends and following them to where you are. WSDM, pages 723–732, 2012.

[16] R. Salakhutdinov and M. Andriy. Probabilistic matrix factorization. NIPS, pages 1257–1264, 2008.

[17] M. Schmidt, A. Niculescu-Mizil, and K. Murphy. Learning graphical model structure using l1-regularization paths. AAAI, pages 1278–1283, 2007.

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

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

[20] M. Ye, P. Yin, and W.-C. Lee. Location recommendation for location-based social networks. GIS, pages 458–461, 2010.

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

[22] V. W. Zheng, Y. Zheng, X. Xie, and Q. Yang. Collaborative location and activity recommendations with gps history data. WWW, pages 1029–1038, 2010.