Space-Time Aware Behavioral Topic Modeling for Microblog Posts

Qiang Qu [†] Cen Chen [‡] Christian S. Jensen [#] Anders Skovsgaard [♮]
[†] Department of Computer Science, Innopolis University
[‡] School of Information Systems, Singapore Management University
[#] Department of Computer Science, Aalborg University
[♮] TrustSkills, Denmark
[†] qu@innopolis.ru [‡] cenchen.2012@phdis.smu.edu.sg [#] csj@cs.aau.dk [♮] anders@trustskills.com

Abstract

How can we automatically identify the topics of microblog posts? This question has received substantial attention in the research community and has led to the development of different topic models, which are mathematically well-founded statistical models that enable the discovery of topics in document collections. Such models can be used for topic analyses according to the interests of user groups, time, geographical locations, or social behavior patterns. The increasing availability of microblog posts with associated users, textual content, timestamps, geo-locations, and user behaviors, offers an opportunity to study space-time dependent behavioral topics. Such a topic is described by a set of words, the distribution of which varies according to the time, geo-location, and behaviors (that capture how a user interacts with other users by using functionality such as reply or re-tweet) of users. This study jointly models user topic interest and behaviors considering both space and time at a fine granularity. We focus on the modeling of microblog posts like Twitter tweets, where the textual content is short, but where associated information in the form of timestamps, geo-locations, and user interactions is available. The model aims to have applications in location inference, link prediction, online social profiling, etc. We report on experiments with tweets that offer insight into the design properties of the papers proposal.

1 Introduction

Microblogging services that enable the posting and browsing of messages containing, e.g., news or local events, are increasingly being used for social interactions.

For example, Twitter has several hundred million active users from around the world who post half a billion messages each day (https://about.twitter.com/company) and is arguably the most important microblogging service. Twitter messages, called tweets, are timestamped and are limited to 140 characters. Twitter supports reply, retweet, and mention functions for tweets, thus enabling social interactions around tweets. We are also witnessing an increased use of geo-enabled mobile devices, most notably smartphones [12]. They offer not only a timely way of using Twitter, but they also offer the ability to associate user location with tweets, yielding geo-tagged tweets.

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

The resulting tweets offer the following information: 1) who posted the tweet; 2) textual content; 3) the time when the tweet was posted; 4) the geo-location from which the tweet was posted; and 5) an associated social behavior (i.e., post, reply, retweet, or mention).

For example, tweet T1 in Table 1 was posted by "@ohcindyoh"; has text that suggest that the tweet concerns a movie; was posted on April 29, 2013; was posted from location @cinema21[1]; and was posted as an original tweet (using "post"). Put differently, tweets may be viewed as being 5-dimensional.

ID	Date	Author	Textual Content
T1	April 29, 2013	@ohcindyoh	Watching Iron Man 3 (with Geng Depo Bangunan at @cinema21).
T2	April 29, 2013	@imabieberchicka	@brailleman89 What are you doing?
T3	April 30, 2013	@imivycaparas	Gorg sis!! Daniel's concert tomorrow): huehuehue im jelly! Buy smth for me!!! Shirt okay): @jiannex

Table 1: Example tweets.

The availability of large collections of such 5-dimensional microblog posts makes it relevant to study an integrated model of social behavioral patterns that exploits all five dimensions. Existing studies have, however, proposed to model topics of social data based on only some of the 5 dimensions [4, 13–15, 18]. These models can be used in applications such as topic mining [3], followee recommendation [4], and location prediction [13]. However, to the best of our knowledge, this study is the first to consider all of the 5 dimensions of microblog posts. More specifically, we consider behaviors that correspond to the social functions offered by the microblogging service, i.e., post, retweet, reply, mention, for tweets. We consider this user behavior together with space and time because all three describe the context in which posts are generated by users.

To exemplify this context, consider again Table 1. Here, tweet T2 is a reply (tweets staring with @"user") from @imabieberchicka to @brailleman89 that concerns their relationship. Some Twitter users use Twitter as a chatting app for interacting with their friends, so that most tweets are associated with the reply behavior. If we model user behaviors and topics jointly as in one study [4], we will find a "reply-daily" topic that concerns mostly daily issues and appears in user replies. By looking at this topic, we can find users that interact with other users using "reply". The distribution of topics can depend on the kind of user behavior (e.g., post, reply), for which reason the topics are called behavioral.

Next, some users use Twitter as a news channel and often retweet news events. For instance, movie fans often talk about new movies, and music fans may often talk about concerts. In this case, we may find that topics are associated with events. For example, tweet T1 is about watching a movie on April 29, 2013. If we observe many tweets talking about the movie "Iron Man" on the same day then there may be an event related to "Iron Man" on that day. It is thus important to consider time information.

Last but not least, some users may be interested only in events happening close to their locations. It is thus beneficial to consider the geo-location of behavioral topics. For example, tweet T3 concerns a concert in the Philippines. It is then likely to be most appropriate to recommend this event to users in the Philippines. This shows that it is also necessary to consider space information.

In sum, it is important to model users, textual content, behaviors, space, and time jointly for microblog posts. We thus propose a space-time dependent behavioral topic model. However, it is difficult to simply aggregate the dimensions of tweets in a regression model as they are of different types.

The proposed modeling has three notable benefits. First, we can identify user groups at similar locations with similar topics during a time period, but with different social behaviors. For example, Twitter users are likely to check-into geographical locations when posting tweets concerning local events. The identification of different

[1]T1 contains a check-in that is regarded as a geo-location tag. If a tweet has no check-in, we use its lat-long as its geo-location.

user behaviors may help us understand their motivations for using Twitter and how actively they interact with the local events. Second, we can profile users and locations according to social behaviors (e.g., reply and retweet behaviors) and the changes of topics over time. Third, we can predict user locations at a specific time given topics and behaviors.

We compare our model with existing models and propose methods for estimating the parameters of the model. Experimental findings from experiments with tweets show that our model is capable of identifying interesting space-time dependent behavioral topics of users and of predicting user locations. The results also suggest that the proposed model is effective for the applications considered.

The rest of the paper is organized as follows. Following a coverage of related work in Section 2, Section 3 presents the proposed model and means of estimating model parameters. The experimental study is presented in Section 4. Section 5 concludes and discusses future work.

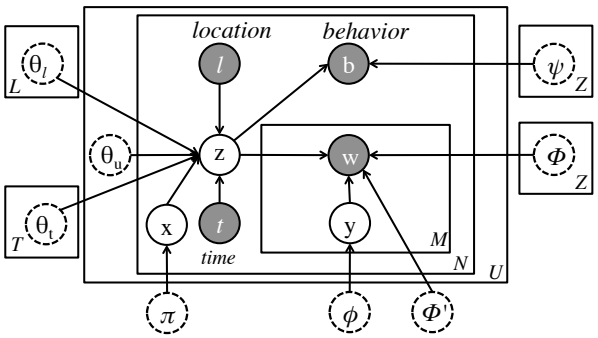
2 Related Work

Recently, geo-tagged and time-stamped social media has drawn much attention [1, 6, 7, 9, 10, 17]. Some studies propose to model topics of microblog posts to understand their social content. Topic models like LDA [2] have been used widely to find hidden "topics" in documents. In these models, each document can be represented in a semantic topic space, which also enables tasks like text classification and document clustering. There is growing interest in adapting topic models to short texts like microblog posts [3, 14, 18].

Twitter-LDA (T-LDA) [3] addresses the shortness of tweets while making two assumptions: 1) one tweet has one hidden topic assignment; and 2) a given tweet may contain both topical words and background words, where the former are words specific to the topic of the tweet and the latter are words that are popular in many tweets. Experiments suggest that T-LDA can capture more meaningful topics than LDA in Twitter data [3], and T-LDA is further extended into Behavior-LDA (B-LDA) [4] to jointly model the topic interests and interactions of a user. B-LDA assumes a universal behavior distribution instead of a personalized behavior distribution for each topic, as the former ensures the behavior information is a property of "topic". In this case, by examining a user's "topic" distribution, one may find personal behavior patterns and topic interests. In other words, a "topic" here is a behavioral topic. To avoid confusion, we refer to a behavioral topic as a *topic* in this study.

One study [13] reviews some of the previous studies that integrate some of the 5 dimensions considered in this paper, and the proposed model (W^4) supports four dimensions (who, when, where, and what). However, W^4 cannot distinguish varying user behaviors. In other words, the model is unable to identify topics from posts by the differences in how the users interact with the content. Moreover, W^4 models time as categorical values consisting of week and weekend days, which is very coarse when aiming to find timely topics. To the best of our knowledge, our study is the first that integrates the 5 dimensions in one model. Further, our model considers location and time at a fine granularity. In experiments, we show results based on the use of fine geographical regions and precise time-stamps of tweets.

Another category of studies relevant to our problem is multi-view clustering [8, 16], where each independent view is able to cluster the data. Generally, the method aims to exploit the multiple views to discover the clusters that agree across the views. For example, the Co-EM algorithm [8] is an expectation maximization algorithm that iteratively performs the expectation step in one view, the result of which is passed to the maximization step in another view. In multi-view clustering, each view corresponds to a representation of the same data with different features, and the goal is to cluster the data by making use of multiple features. Our problem is not a clustering problem, and it has a different goal than unsupervised clustering.



l	a location identifier $1 \leq l \leq L$
b	a behavior in $\mathcal{B} = \{post, retweet, reply, mention\}$
z	a topic identifier $1 \leq z \leq Z$
u	a user identifier $1 \leq u \leq U$
n	a tweet identifier $1 \leq n \leq N$
w	a word identifier $1 \leq w \leq M$
t	a time-stamp identifier $1 \leq t \leq T$
x, y	switches, $x \in [0, 2]$, $y \in [0, 1]$
ψ_z	topic-specific behavior distribution
ϕ_z	topic-specific word distribution
ϕ'	background word distribution
θ_u, θ_l, θ_t	user, location, and time dependent topic distributions
α_u, α_l, α_t	priors of θ_u, θ_l, and θ_t
π, φ	Bernoulli and multinomial distributions

Figure 1: Plate notation for our space-time dependent behavioral topic model for microblog posts. The dashed variables will be collapsed out during Gibbs sampling [11]. Priors over all the multinomial or binomial distributions are omitted for clarity.

3 Model

In this section, we present our space-time dependent behavioral topic model as shown in Figure 1.

3.1 Space-Time Dependent Behavioral Topic Model

In B-LDA [4], it is assumed that all tweets posted by a user concern the user's own interests. However, in many cases, users will not only post tweets according to their own topics of interest, but may also post tweets that concern temporal events and location-dependent topics. As a result, at least two additional dimensions may be built into the model, i.e., *space* and *time*. We propose a probabilistic model that jointly models the space and time of tweets for behavior-topic analysis. In the space-time dependent behavior-topic model, we assume to have three types of topic distributions, i.e., user-dependent, space-dependent, and time-dependent topic distributions. Below we present the full model that considers 5 dimensions: user, text, time, location, and behavior.

We assume to have a data set that contains U users. A user u has N_u ($1 \leq u \leq U$) tweets. We use $M_{u,n}$ ($1 \leq n \leq N_u$) to denote the number of words in nth tweet of uth user, and $w_{u,n,m}$ ($1 \leq m \leq M_{u,n}$) to denote the mth word in nth tweet of uth user, where $1 \leq w_{u,n,l} \leq V$ and V is the vocabulary size. Next, $l_{u,n}$ and $t_{u,n}$ denote the location and time, respectively, of the nth tweet of the uth user. Similar to B-LDA, our model assumes a space $\mathcal{B}$ containing all possible types of behaviors. In the case of Twitter, $\mathcal{B} = \{post, retweet, reply, mention\}$. We use $b_{u,n} \in \mathcal{B}$ to denote the behavior of the nth tweet of the uth user.

We now present our model. First, we assume that there are Z hidden topics, where each topic has a multinomial word distribution ϕ_z and a multinomial behavior distribution ψ_z. We pose Dirichlet priors η and β on ϕ_z and ψ_z, respectively.

$$\forall z(\phi_z \sim \text{Dir}(\beta) \text{ and } \psi_z \sim \text{Dir}(\eta)) \tag{31}$$

Recall that we assume to have three types of topic distributions, i.e., a user-dependent distribution θ_u, a space-dependent distribution θ_l, and a time-dependent distribution θ_t. Similarly, we pose Dirichlet priors $\alpha_u, \alpha_l, \alpha_t$ on these distributions.

$$\forall u(\theta_u \sim \text{Dir}(\alpha_u)), \forall l(\theta_l \sim \text{Dir}(\alpha_l)), \text{ and } \forall t(\theta_t \sim \text{Dir}(\alpha_t)) \tag{32}$$

Each tweet has a single topic that is sampled from one of the three topic distributions θ_u, θ_l, and θ_t. Let $\text{Multi}(\pi) \sim \text{Dir}(\gamma)$. We then use a switch $x_{u,n} \sim \text{Multi}(\pi)$ to choose a topic from the three distributions (values

0, 1, and 2 of $x_{u,n}$ indicate switches of the user, location, and time dependent distributions).

$$z_{u,n} \sim \begin{cases} \text{Multi}(\theta_u) & \text{if } x_{u,n} = 0 \\ \text{Multi}(\theta_l) & \text{if } x_{u,n} = 1 \\ \text{Multi}(\theta_t) & \text{if } x_{u,n} = 2 \end{cases}$$

For a tweet with a topic label $z_{u,n}$ ($1 \leq z_{u,n} \leq Z$), the words in this tweet are generated from two multinomial distributions, namely a background word distribution ϕ' and a topic specific word distribution ϕ. Similarly, they are with Dirichlet priors $\phi' \sim \text{Dir}(\beta')$ and $\phi \sim \text{Dir}(\beta)$. Let $\text{Multi}(\varphi) \sim \text{Dir}(\rho)$. We then use a switch $y_{u,n,m} \sim \text{Multi}(\varphi)$ to choose a word from the two distributions indicated by 0 and 1 values of $y_{u,n,m}$.

$$w_{u,n,m} \sim \begin{cases} \text{Multi}(\phi') & \text{if } y_{u,n,m} = 0 \\ \text{Multi}(\phi_{z_{u,n}}) & \text{if } y_{u,n,m} = 1 \end{cases}$$

3.2 Learning and Parameter Estimation

We use the Collapsed Gibbs sampler [11] to obtain samples of the hidden variable assignments and to estimate the model parameters from these samples. We show the derived Gibbs sampling formulas in the following. Proofs are similar to those given in related work [4].

For tweet n of user u, we jointly sample a switch $x_{u,n}$ and its topic label $z_{u,n}$.

$$\begin{aligned} &p(z_{u,n} = z, x_{u,n} = x \mid \mathbf{Z}_{\neg_{u,n}}, \mathbf{X}, \mathbf{L}, \mathbf{T}, \mathbf{B}) \\ &= \frac{n^x + \gamma}{\sum_{x' \in [0,2]}(n^{x'} + \gamma)} \cdot \frac{n^b_z + \eta}{\sum_{z'}(n^b_z + \eta)} \cdot \left[\frac{n^z_u + \alpha_u}{\sum_{z'}(n^{z'}_u + \alpha_u)}\right]^{x=0} \cdot \left[\frac{n^z_l + \alpha_l}{\sum_{z'}(n^{z'}_l + \alpha_l)}\right]^{x=1} \cdot \left[\frac{n^z_t + \alpha_t}{\sum_{z'}(n^{z'}_t + \alpha_t)}\right]^{x=2}, \end{aligned} \tag{33}$$

where l, t, and b denote the location, time, and behavior information; $n^{z'}_u$ refers to the number of times topic z' co-occurring with user u; and other ns are defined in the same way.

For each word $w_{u,n,m} = w$ in tweet n of user u, we sample its switch $y_{u,n,m}$ as follows.

$$p(y_{u,n,m} = y \mid \mathbf{Y}_{\neg_{u,n}}, \mathbf{X}, \mathbf{Z}, \mathbf{L}, \mathbf{T}) = \frac{n^y + \rho}{\sum_{y' \in [0,1]}(n^{y'} + \rho)} \cdot \left[\frac{n^w_{y=0} + \beta}{\sum_{w'}(n^{w'}_{y=0} + \beta)}\right]^{y=0} \cdot \left[\frac{n^w_z + \beta}{\sum_{z'}(n^w_{z'} + \beta)}\right]^{y=1}, \tag{34}$$

where $n^w_{y=0}$ refers to the number of times word w being labeled as a background word.

With the Collapsed Gibbs sampler, we can make the following estimation of the model parameters:

$$\theta_{u,z} = \frac{n^z_u + \alpha_u}{\sum_{z'} n^{z'}_u + Z\alpha_u} \quad \text{user-topic distribution} \tag{35}$$

$$\theta_{l,z} = \frac{n^z_l + \alpha_l}{\sum_{z'} n^{z'}_l + Z\alpha_l} \quad \text{location-topic distribution} \tag{36}$$

$$\theta_{t,z} = \frac{n^z_t + \alpha_t}{\sum_{z'} n^{z'}_t + Z\alpha_t} \quad \text{time-topic distribution} \tag{37}$$

$$\psi_{z,b} = \frac{n^b_z + \eta}{\sum_{b'} n^{b'}_z + B\eta} \quad \text{topic-behavior distribution} \tag{38}$$

$$\phi_{z,w} = \frac{n^w_{z,y=1} + \beta}{\sum_{w'} n^{w'}_{t,y=1} + V\beta}, \quad \text{topic-word distribution} \tag{39}$$

where n^z_u is the number of times z is sampled for user u and n^b_z is the number of times behavior b co-occurs with topic z.

4 Experimental Study

We proceed to evaluate the proposed model. We first describe the datasets and then present the experimental setup. Finally, we report on findings of a set of experiments.

4.1 Data and Settings

We collected all world-wide geo-tagged tweets from the public Twitter Streaming API from April 29 to July 2, 2013, and we choose 10,000 users at random and use all their tweets. We further select 90% of all the tweets at random for training our model and use the remaining 10% of all tweets for evaluating our model.

Our model is able to find user-specific, space and time dependent behavioral topics, making it useful for several real-world tasks. To evaluate the model, we

1. qualitatively analyze the learned word distributions and topic distributions from the model, and we
2. quantitatively evaluate the model against baseline models for the task of location prediction.

In this study, we focus on location and time relevant topics. Our model inherits its behavior dimension from B-LDA. We thus do not discuss behavioral topics.

We ran 1000 iterations of Monte Carlo EM. For the Gibbs sampling steps, we ran 400 iterations for burn-in, and we sampled every 10 iterations to reduce auto-correlation. We fixed the number of topics at 20. (We varied this number from 10 to 100 with a step size of 10 and found the resulting topics to be most meaningful at around 20 by manual examination). For our models and competing baselines, we use grid search on a development set to select the model parameters.

4.2 Qualitative Analysis

(Topics.) Table 2 presents top topic-specific words for some sample topics. The experimental findings show that Twitter users often talk about themselves, for example, topic "daily life" is a popular topic that mostly concerns the users' daily updates. Similarly, topic "school" looks to be on updates about school. The topic "music" is about songs, country music, pandora, etc. All these topics are readily identified based on their top topical words. They can also serve as interpretable labels for the corresponding tweets or users.

We note that some of the extracted topics are featured with location information. For example, tweets related to topic "movie" are mostly posted from locations close to a cinema. This suggests that some locations have their own topics and relevant words; thus, based on the words used by users, we can draw clues about users' locations. In light of this, we study location prediction in Section 4.3.

(Location and Time Dependent Events.) Unlike related work [13], the proposed model considers temporal information at a fine granularity. This allows us to discover bursty events, i.e., topics with a sudden increase of usage. We define a burstiness score of topic t on day d as $s(t,d) = \frac{c_{t,d} - c_{t,d-1} + 1}{c_{t,d-1} + 1}$, where $c_{t,d}$ denotes the number of tweets with topic t on day d.

Table 3 visualizes top bursty topics sorted by $s(t,d)$ as obtained using our model. We find that all these bursty events are meaningful. The first bursty event is about the release of the movie Iron Man 3. The second concerns a concert. The third one concerns a political event. Note that in the proposed model, each topic has a location distribution; Thus, all the bursty events above have a location dimension. Close examination shows that the first bursty event has tweets are from all over the world and is global. The second one is more localized as its tweets are from the Philippines. The third one happened in the UK when the UKIP leader Nigel Farage was on a campaign visit to Edinburgh. By using our model, we find that the locations associated with the event are indeed from the UK. In all, we find that by considering space and time in the modeling of topics of microblog posts, we can obtain better insights into the behavioral topics of users, locations, and times.

"daily life"	"god"	"cars"	"school"	"music"	"movie"	"food"	"drink"
good	god	car	school	song	movie	food	drink
today	lord	drive	year	music	watch	eat	smoke
tired	world	ride	class	shows	watching	ice	water
early	jesus	house	summer	trend	show	cream	beer
nap	bless	hit	days	listen	funny	pizza	drinking
ready	live	driving	hate	listening	fast	chicken	bottle
day	man	street	test	album	movies	breakfast	smoking
school	beautiful	walking	exam	favorite	game	chocolate	blunt
wake	give	walk	back	world	favorite	hot	cold
shower	woman	bus	start	songs	guy	cheese	juice
long	good	road	ill	love	episode	ate	drunk
feeling	blessed	hate	homework	country	purge	hungry	coffee
woke	love	work	final	pandora	wolf	dinner	smell
night	life	gas	math	topic	teen	cake	cup
awake	pray	truck	english	taylor	family	ill	drank

Table 2: Top topic-specific words from $\phi_{z,w}$ for sample topics. Labels are assigned manually.

Dates	Tweets	Label
April 29, 2013	Iron man 3 wiff @rasekarini (@ Studio 21 - @cinema21 w/ 18 others) Uuurgh Can't wait to watch "IronMan", Seems like it's awesome movie Hype for this new Iron Man movie....I ♡ Marvel Watching Iron Man 3 (at @cinema21) http://t.co/QMVlvpkgIk	Iron Man 3
April 30, 2013	Daniel padilla live paperview yipee < 3 Okaaay so like, naa man daw payperview sa concert ni daniel padilla =)) Daniel padilla invades not just araneta,but also the twitter world.. Rocking my souvenir!! Daniel padilla concert #DanielLiveAtTheBigdome	Daniel Live! Concert
May 16, 2013	Yes. Yes, he is. RT @juliahobsbawm: Nigel Farage is a Black Swan. Nigel Farage has a great taste in suits and hats it must be said! Well that's my vote. Viva Nigel Farage! http://t.co/MC7k84YgIO Johnny don't do Nigel Farage as he would look exactly the same #UKIP	Nigel Farage is heckled

Table 3: Bursty topics found by our model and sample tweets. Labels are assigned manually.

4.3 Location Prediction

We apply our model to a location prediction task. Specifically, given a tweet from a user, the task is to predict the location where the user posted it. The intuition is that many locations have their own topics. For example, if a location is a food court, people tend to tweet more about food in this location. Our method is that we first obtain location-dependent topics learned by the proposed model; then, given a tweet with a set of words and a behavior, we estimate its topics and find the most relevant location, detailed as follows.

For tweet n from user u with words $w_{u,n}$ and behavior $b_{u,n}$, we predict its location by using this formula: $l_{u,n} = \text{argmax}_l\, p(l|w_{u,n}, b_{u,n}, \psi, \phi, \phi', \theta)$. Here, ψ, ϕ, ϕ', and θ are learned using Equations 36-39. We further

compute $p(l|w_{u,n}, b_{u,n}, \psi, \phi, \phi', \theta)$ as follows.

$$\begin{aligned} p(l|w_{u,n}, b_{u,n}, \psi, \phi, \phi', \theta) &\propto p(l)p(w_{u,n}, b_{u,n}|l) \\ &= p(l)\sum_z p(z|l)p(w_{u,n}, b_{u,n}|z) \\ &= p(l)\sum_z \theta_{l,z}p(w_{u,n}|z)p(b_{u,n}|z) \\ &= p(l)\sum_z \theta_{l,z}\psi_{z,b_{u,n}}\prod_{w\in w_{u,n}}\phi_{z,w} \end{aligned} \tag{40}$$

For simplicity, we assume all the words are topic-specific, and we approximate $p(l)$ by using the popularity of location l.

Recall that we use 90% of all tweets for learning and that we have held out 10% of all tweets for testing. For each tweet in the test data, we compute its probability of belonging to a certain location l using the above method. We then sort the locations based on the probabilities. The higher the real location of the tweet is ranked, the better our method is. We consider three baseline methods:

1. Random. By using random guessing, the expected ranking of the real locations will average at around 50%.

2. Majority. The majority baseline always ranks the locations by their popularity. This works well when the held-out tweet locations are from popular locations.

3. Clustering method. This method treats all tweets with the same location as a cluster, and for a new tweet, we compute its similarity to all the clusters and rank all the locations according to the similarity scores. To measure the similarity between a tweet and a cluster of tweets, we use the averaged Jaccard index score.

As for evaluation, we use these metrics: average ranking r_{average} (the lower the better), median ranking r_{median} (the lower the better), and mean reciprocal rank MRR (the higher the better), defined as follows. $r_{\text{average}} = \frac{1}{|D_{test}|}\sum_{d\in D_{test}}\frac{r_d}{|L|}$, $MRR = \frac{1}{|D_{test}|}\sum_{d\in D_{test}}\frac{1}{r_d}$. Metric r_{median} is similar to r_{average}, but uses the median ranking instead of average. Here, r_d refers to the real location's ranking for tweet d, $|L|$ is the total number of locations, and D_{test} is the test set. These criteria have also been used for a similar task, followee recommendation [4, 19].

Table 4 shows the results. The majority method performs worse than the random method. This means that the held-out tweet locations are often from less frequent locations. By Wilcoxon signed-rank test [5] and the results in Table 4, we obtain that the clustering method outperforms both majority and random methods at 0.1% significance level. This implies that many locations indeed have their own location-specific behavioral topics. Our method also outperforms the other methods at 0.1% significance level by Wilcoxon signed-rank test. Using average ranking, our method ranks the real locations in the top 12.6% of all the locations, and with a median ranking at around 2.7% that means that the real location of a given tweet is ranked in the top 2.7% of all the locations. Since the location set size is large in our data set, the findings show that our method can learn good location specific topics and topic specific words.

Metric	Our Model	Random	Majority	Clustering
r_{average}	**0.126**	0.5	0.55	0.132
r_{median}	**0.027**	0.5	0.54	0.103
MRR	**0.090**	0.0001	0.0003	0.015

Table 4: Comparison of the methods used for location prediction.

A given tweet may not necessarily be location-specific but could be user-specific or time-specific. To address this, we propose a simple way to compute a confidence score to measure whether a tweet is location-specific or not. We define the confidence score $s(d)$ as the aggregated probability of the tweet d belonging to a certain location: $s(d) = \sum_l p(l|w_d, b_d, \psi, \phi, \phi', \theta)$. Below we show the results of our model in terms of different confidence scores.

Metric	10%	20%	50%	80%	100%
r_{average}	**0.076**	0.104	0.126	0.130	0.126
r_{median}	0.034	0.036	0.039	0.034	**0.027**
MRR	0.073	0.071	0.070	0.081	**0.090**

Table 5: Findings for location prediction. We use set of tweets with different confidence scores. Thus, n% means that the tweets with the n% highest confidence scores are used.

Using r_{average}, our method is best at tweets with top 10% confidence scores. Our method ranks the real locations in the top 7.6% when using the top 10% most confident tweets, while our method ranks the real locations in the top 12.6% when considering all tweets. Results at 10% are better than at 20% and 50% in terms of both r_{median} and MRR, which indicates that top confident tweets often benefit location prediction. The table also shows that the MRR score at 10% is not as high as at 100%. The reason may be that on the tweets with top 80–100% confidence scores, the variance is smaller than that using tweets with top 10–50% confidence scores. Similar observation can be found for r_{median}, and the results also show that r_{median} seems to be rather insensitive to the percentage.

5 Conclusion and Future Work

In this study, we propose to model space and time dependent behavioral topics of microblog posts that associated with text, timestamps, geographical locations, and user behaviors with users. The experiments on Twitter data demonstrate our model is able to identify useful and insightful user behavioral topics with a fine spatial and temporal granularity.

There may exist a range of applications of our model, including user location inference, link prediction, user or location profiling by the changes of topics over time, burst event detection, and automatic tagging semantic text to geographical locations. Applications such as these deserve exploration in future work. Moreover, it may also be promising to integrate other contextual types, such as popularity of images on Instagram, in our model, or to find a generalized way to integrate social context with textual content in the model.

Acknowledgments

We thank Minghui Qiu and Anna Tigunova for the contribution in parts of the discussion and programming. This research is supported by the Russian Science Foundation under Grant No. 15-11-10032.

References

[1] X. Cao, L. Chen, G. Cong, C. S. Jensen, Q. Qu, A. Skovsgaard, D. Wu, and M. L. Yiu. *Spatial Keyword Querying*. In ER, 2012, pp. 16–29.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. *Latent dirichlet allocation*. J. Mach. Learn. Res., vol. 3, pp. 993–1022, 2003.

[3] X. Zhao, J. Jiang, J. Weng, J. He, E. P. Lim, H. Yan, X. Li. *Comparing Twitter and Traditional Media Using Topic Models*. In ECIR, 2011, pp. 338–349.

[4] M. Qiu, F. Zhu, and J. Jiang. *It Is Not Just What We Say, But How We Say Them: LDA-based Behavior-Topic Model*. In SDM, 2013, pp. 794–802.

[5] I. C. A. Oyeka and G. U. Ebuh. *Modified Wilcoxon Signed-Rank Test*. Open Journal of Statistics, no. 2, pp. 172–176, 2012.

[6] Q. Qu, S. Liu, B. Yang, and C. S. Jensen. *Integrating non-spatial preferences into spatial location queries*. In SSDBM, 2014, Article 8.

[7] C. R. Vicente, D. Freni, C. Bettini, and C. S. Jensen. *Location-Related Privacy in Geo-Social Networks*. IEEE Internet Computing, vol. 15, no. 3, pp. 20–27, 2001.

[8] S. Bickel and T. Scheffer. *Multi-view clustering*. In ICDM, 2004, pp. 19–26.

[9] Y. Zhou, G. Cong, B. Cui, C. S. Jensen, and J. Yao. *Routing Questions to the Right Users in Online Communities*. In ICDE, 2009, pp. 700–711.

[10] Q. Qu, S. Liu, B. Yang, and C. S. Jensen. *Efficient Top-k Spatial Locality Search for Co-located Spatial Web Objects*. In MDM, 2014, pp. 269–278.

[11] J. S. Liu, *The Collapsed Gibbs Sampler in Bayesian Computations with Applications to a Gene Regulation Problem*. Journal of the American Statistical Association, vol. 89, no.427, pp. 958–966.

[12] C. Hage, C. S. Jensen, T. B. Pedersen, L. Speicys, and I. Timko. *Integrated data management for mobile services in the real world*. In VLDB, 2003, pp. 1019–1030.

[13] Q. Yuan, G. Cong, Z. Ma, A. Sun, and N. Magnenat-Thalmann. *Who, where, when and what: discover spatio-temporal topics for twitter users*. In SIGKDD, 2013, pp. 605–613.

[14] Y. Wang, E. Agichtein, and M. Benzi. *TM-LDA: efficient online modeling of latent topic transitions in social media*. In SIGKDD, 2012, pp. 123–131.

[15] M. C. Gonzlclez, C. A. Hidalgo, and A. L. Barabasi. *Understanding individual human mobility patterns*. Nature 453, pp. 479–482, 2008.

[16] S. Sun. *A survey of multi-view machine learning*. Neural Computing and Applications, vol. 23, no. 7–8, pp. 2031–2038, 2013.

[17] Q. Qu, S. Liu, C. S. Jensen, F. Zhu, and C. Faloutsos. *Interestingness-Driven Diffusion Process Summarization in Dynamic Networks*. In ECML/PKDD, 2014, pp. 597–613.

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

[19] W.-Y. Chen, J.-C. Chu, J. Luan, H. Bai, Y. Wang, and E. Y. Chang, *Collaborative filtering for orkut communities: discovery of user latent behavior,* in WWW, 2009, pp. 681–690.