A Local-Global LDA Model for Discovering Geographical Topics from Social Media

Siwei Qiang, Yongkun Wang, and Yaohui Jin

Shanghai Jiao Tong University Network and Information Center, Shanghai
{qiangsiwei,jinyh,ykw}@sjtu.edu.cn

Abstract. Micro-blogging services can track users' geo-locations when users check-in their places or use geo-tagging which implicitly reveals locations. This "geo tracking" can help to find topics triggered by certain events in certain regions. However, discovering such topics is very challenging because of the large amount of noisy messages (e.g. daily conversations). This paper proposes a method to model geographical topics, which can filter out irrelevant words by different weights in the local and global contexts. Our method is based on the Latent Dirichlet Allocation (LDA) model but each word is generated from either a local or a global topic distribution by its generation probabilities. We evaluated our model with data collected from Weibo, which is currently the most popular micro-blogging service for Chinese. The evaluation results demonstrate that our method outperforms other baseline methods in several metrics such as model perplexity, two kinds of entropies and KL-divergence of discovered topics.

Keywords: Geolocation, Geographical topics; Topic modeling; Latent Dirichlet Allocation

1 Introduction

Micro-blogging services including Twitter and Weibo have emerged as a medium in spotlight for online users to share breaking news or interesting stories in their lives and update their status anywhere and anytime in their daily lives. With the advancement of positioning technology, the popularity of low-cost GPS chips and wide availability of smart phones, large-scale crowd-generated social media data with geographical records have become prevalent on the web and can also be easily collected. Such textual data with geo-coordinates or geo-tagged locations usually contain landmark information (e.g., scenic spots or famous restaurants) or information on local events (e.g., movies, vocal concerts, exhibitions or sports games), and hence, can provide us rich and interpretable semantics on different locations. It is also possible to infer inherent geographic variability of topics across various locations.

In recent years, a significant amount of research have been conducted on addressing the questions of how the information is created and shared in different geographic locations and how the spatial and linguistic characteristics of people

vary across regions. Among them, a considerable amount of studies have been conducted on GPS-associated documents including organizing geo-tagged documents or photos and studying user movement for POI recommendation, user prediction and time prediction. They try to address the following two needs. The first is to discover different topics of interests those are coherent in geographical regions. For example, a city is usually formed by different functional sub-regions, such as business area, residential area and entertainment area. The second is to comparing the topics discovered across different geographical locations. For example, people would like to know where is the landmarks of the city for tourists or which places they could go when they plan to go shopping or have fun during the weekends.

However, the challenge is that the messages with geo-locations are mixed with overwhelming noisy messages of daily chats or expressions of personal emotions, which have little or no relations to the location context. For example, in the city of Shanghai, *Waitan* is a famous waterfront and one of the most popular scenic spots for tourists. However, even in such a spot, Weibo are still full of daily conversations and greetings such as 'Good night' or 'Have a nice weekend', which has no local semantics. When taking all local posts into account, the meaningfulness or concentrativeness of the discovered topics can be compromised. Therefore, it is very difficult to discover meaningful geo-location topics by existing methods such as inferring occurrences of words from local posts.

This paper proposes an effective method to handle noisy messages and model geographical topics of different locations. The proposed method is based on the Latent Dirichlet Allocation (LDA) [2] topic model. The intuitive idea is that, for a noise specific location, the words used by users are different between **a)** daily conversations and **b)** the description of the landmark or local events. The former is relatively consistent across different locations, and denoted as *global context*, while the latter, which is essentially helpful to identify the true characteristics of the region or area, varies by sites and are denoted as *local context*. Our method takes the local and global contexts into consideration, and different from all existing models to reveal spatial topics, it models each word to be generated from either its local or the global topic distribution by its estimated probabilities. The proposed strategy is able to distinguish locally featured words from noise and improve the quality of discovered topics.

Our evaluation based on two typical social media datasets. One is from Weibo, which is a Chinese micro-blogging website. Akin to a hybrid of Twitter and Facebook, it is one of the most popular sites in China, in use by well over 30% of Internet users, with a market penetration similar to the United States' Twitter. The other is from Yelp, which publish crowd-sourced reviews about local businesses, as well as the online reservation service and online food-delivery service. Our model is evaluated with several metrics widely used in assessing topic models, such as perplexity and KL-divergence, together two kinds of entropies, topic entropy and location entropy to assess the concentrativeness of the discovered topics. The evaluation results demonstrate that our method outper-

forms other baseline methods and show its superiority in information filtering and geographical topic discovery.

2 Related Work

In this section we discuss some work related to our study, including geo-tagged social media mining, topic modeling and local word detection.

Mobility or posting pattern mining with geo-tagged social media data become a hot topic with the development of GPS technology. The activities of mobile users are typically represented as follows: a user appears at a certain location (with a pair of latitude and longitude coordinates), and leaves a post (e.g., Weibo or review), which is likely semantically related to the user and/or the location [18]. The mining problem is usually formulated as finding various mobility or posting patterns from user activities, such as frequent patterns, periodic behaviors, representative behaviors and activity recognition [3, 7, 13]. In literature, numerous methods have been proposed to extract such patterns from the social media data. Representative works include stop and move detection, significant place extraction, frequent regular pattern discovery, transportation mode recognition. However, these works mainly focus on the trajectory or posting patterns, and seldom explore the contextual semantics of user-generated contents.

Topic modeling is a classic task to enable text analysis at a semantic level and to discover hidden semantic structures in a text body. The most representative and widely used topic models are probabilistic latent semantic analysis(pLSA) [1] and LDA [2]. Both are generative statistical models, and assume that in a given dataset each document is associated with a topic distribution, and each topic with a word distribution, the difference is that in LDA, the topic distribution is assumed to have a Dirichlet prior, and in practice, this results in more reasonable mixtures of topics in a document. Recently, in order to support location-aware information retrieval or to compare topics across geographical locations, there are many works in the area of geographical topic modeling [4,5,8,9,12,14–17,19, 21–23]. For example, Yin et al. proposes and compares three ways of modeling geographical topics, including a location-driven model, a text-driven model, and a joint model called LGTA [14], which combines geographical clustering and topic modeling into one framework. In this model, the coordinates in each document are drawn from a 2D Gaussian distribution and the region is drawn from a Multinomial distribution over all regions. Hong et al. models diversity in tweets based on topical diversity, geographical diversity, and an interest distribution of the user [16]. Further, it takes the Markovian nature of users' locations into account and identifies topics based on location and language. The spatial Topic (ST) Model for location recommendation has been proposed by Hu and Ester recently to capture the correlation between users' movements and between user interests and the function of locations [18]. A hierarchical topic model which models regional variations of topics has been presented by Ahmed et al., which combines distributions over locations, topics, and over user characteristics, both in terms of location and in terms of their content preferences [20]. Unlike previous

work, it automatically infers both the hierarchical structure over content and over the size and position of geographical locations, and gains higher accuracy on location estimation from Tweets. Although all the above works discover regions and geographical topics, they do not consider the overwhelming noisy messages in user-generated contents, which can have a major impact on the results.

Another line relevant to our research is local word detection. The general idea is that when a location specific event of interest takes place, there can be a surge in the volume of documents related to the event, and as a result, such a surge of information can be utilized to identify location-characterized topics. Based on the premise that local words should have concentrated spatial distributions around their location centers, Backstrom et al. proposes a spatial variation model for analyzing geographic distribution of terms in search engine query logs [6], and this method has been used by Cheng et al. to decide whether a word is local or not [12]. Meanwhile, Mathioudakis et al. uses spatial discrepancy to detect spatial bursts, which identifies geographically focused information bursts, attribute them to demographic factors and identify sets of descriptive keywords [10]. In these work, whether the word is local or not is determined by a assigned locality score. However, it is demonstrated by Wu et al. that this method can be erroneous since it assumes one peak density distribution while many local words can have multiple peaks [24]. Our work is different from the existing works in that, word locality is not generated directly but evaluated by the generation probability and is not associated with a static locality score, which means that a word (e.g. car) can be both *non-local* for a majority of locations and also *local* for a few particular locations (e.g. automobile 4S shops).

3 Method

3.1 Local-Global LDA Model

In this section, we propose a novel topic model for geo-tagged social media texts called LGLDA (Local-Global LDA Model), which combines noise filtering and topic modeling into one framework. To begin with, we define the notations used in this paper as listed in Table 1.

To discover geographical topics, the spatial structure of words should be encoded. The words that are close in space are likely to be clustered into the same geographical topic. However, in our dataset, the geographical distance of two words cannot be calculated due to the loss of the exact geo coordinations, but each peace of text is associated with a location tag, therefore if two words come from texts with the same location tag, they are close, otherwise they are distant. Furthermore, if the exact geo coordinations are assessable, the closeness of any two words can be calculated by Euclidean distance, and our model can be modified by assuming that geographical distribution of each region follows a Gaussian distribution. Hence, the words that are close in space are more likely to belong to the same region, so they are more likely to be clustered into the same topic.

Table 1. Notations used in the paper.

Notation	Description
θ_g	Global topic distribution
θ_l	Local topic distribution for location l
ϕ	Word distribution for topic k
ω	Location relevence
z_e	Latent topic
w	Observed word
α_l	Multinomial distribution prior for θ_l
α_g	Multinomial distribution prior for θ_g
β	Multinomial distribution prior for ϕ
γ	Binomial distribution prior for ω
L	Number of locations
D_l	Number of documents in location l
N_d	Number of words in document d

In our scenario, each geo-located document d is tagged with a location l, and contains a set of words w_d. A geographical topic z is a meaningful theme shared by similar locations, and each location is associated with a topic distribution $p(z|l)$.

We formalize our model based on the following intuitions. Firstly, words close in space are likely to be clustered into the same geographical topic. Therefore, topics are generated from locations instead of individual documents. Secondly, locally featured words have a more compact geographical scope. For example, 'bravo' is a word for a performer, so that it is more possible to be used at a theater, a concert or a stadium rather than other places. On the contrary, noisy words (e.g., happy, love, city) can have a much wider spatial range. However, some words could be local for certain locations although these words are commonly used at many places. In our method, the role (local or non-local) of a word is determined by its generation probabilities of its local and global semantic contexts. Therefore, our model is named as Local-Global LDA model (or LGLDA).

The graphical representation of our model is shown in Figure 1. Shaded nodes indicate observed variables or priors, while light ones represent latent variables. In order to keep a small set of parameters for simplification, in our model there are one shared set of topics with two different distributions θ_l and θ_g for the local topics and the global topics, respectively. It might be interesting and reasonable to utilize two kinds of ϕ for words' local and global distributions corresponding to the two topic distributions, and we would like to study it in our future work.

For a collection of L locations, geo-tagged by D documents, each contains N words, the topic of each word can either be drawn from θ_l or from θ_g. Topic assignment is denoted by z_e, and $e(= l/g)$ indicates whether it is drawn from the local or the global. Since micro-blog is length limited, it is likely to have focused concept. Therefore, if a document is location relevant, each word in it is more likely to be relevant. This relevance of each document is indicated by ω,

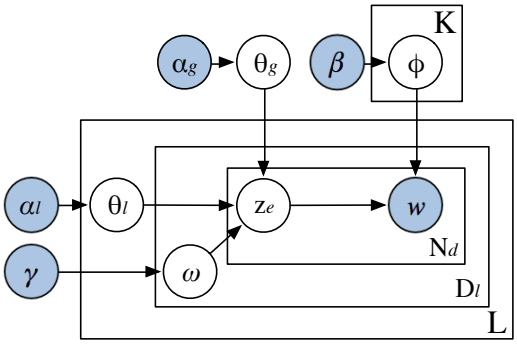

Fig. 1. Graphical representation of the proposed Local-Global LDA model.

with binomial distribution prior γ. Finally, the word distribution in K topics is denoted by ϕ. α_l, α_g, β are priors for θ_l, θ_g and ϕ, respectively.

In order to weight between local and global distributions, we add an additional parameter, named local-global weight ratio. Assume the local and global topic distributions are $\theta_l = [p_{l,1}, ..., p_{l,K}]$ and $\theta_g = [p_{g,1}, ..., p_{g,K}]$ respectively, topic assignment is drawn from a concatenated distribution θ in Equation (1).

$$\begin{aligned} \theta &= \frac{\lambda}{\lambda+1}\theta_l p(e=l|w) \oplus \frac{1}{\lambda+1}\theta_g p(e=g|w) \\ &= [\frac{\lambda p_{l,1} p_w^{(l)}}{\lambda+1}, ..., \frac{\lambda p_{l,K} p_w^{(l)}}{\lambda+1}, \frac{p_{g,1} p_w^{(g)}}{\lambda+1}, ..., \frac{p_{g,K} p_w^{(g)}}{\lambda+1}] \end{aligned} \tag{1}$$

When λ is too large, the global word set is narrowed and ineffective for noise filtering, while when λ is too small, the size of local words is sparse and it fails to discover meaningful topics. Therefore, an appropriate λ is crucial. In our experiment, it is optimized by estimating the model's perplexity (as illustrated in Figure 2 in Section 4.4).

The generative process of our model is summarized in Algorithm 1.

3.2 Model Inference

Like most Bayesian models, collapsed Gibbs sampling was used for model inference. We present the conditional probability of its latent variables $z_e, \theta_l, \theta_g, \phi$ and w for sampling. Details are omitted for limited space. It is assumed that topic distributions θ_l, θ_g and word distribution ϕ of each topic k are drawn from dirichlet distributions of their respective priors α_l, α_g and β, while locality relevance ω are drawn from a binomial distribution with prior γ.

The conditional probability for sampling the topic assignment z_e of each word is computed in Equation (2), where $z_{e,i} = k$ and $e_i = \kappa$ represent the

Algorithm 1 Generative process of LGLDA model.

for each l-th location **do**
  Draw a Dirichlet distribution over all latent topics $\theta_l \sim Dirichlet(\alpha_l)$.
end for
Draw a Dirichlet distribution over all latent topics $\theta_g \sim Dirichlet(\alpha_g)$.
for each k-th topic **do**
  Draw a Dirichlet distribution over all words $\phi \sim Dirichlet(\beta)$.
end for
for each l-th location **do**
  for each d-th Document **do**
    Draw a Bernoulli distribution $\omega \sim Dirichlet(\gamma)$.
    for each w-th word position **do**
      Draw a topic from Multinominal distribution $z_e \sim Dirichlet(\theta_{l|g})$.
      Draw a word from Multinominal distribution $w \sim Dirichlet(\phi)$.
    end for
  end for
end for

assignments of the ith word to topic k, and mark the word as local if $\kappa = 1$ otherwise non-local. $\lambda_\kappa = \frac{\lambda}{\lambda+1}(\kappa = l)$ or $\frac{1}{\lambda+1}(\kappa = g)$. $n_{-i,\kappa,k}$ represents the word count with locality assignment κ and topic assignment k, and $-i$ means not including the ith word.

$$p(z_{e,i} = k, e_i = \kappa | z_{-i}, e_{-i}) \propto \lambda_\kappa \cdot \frac{n^{(d_i)}_{-i,\kappa,k} + \gamma_\kappa}{n^{(d_i)}_{-i,\cdot,k} + \gamma_l + \gamma_g} \cdot \frac{n^{(l_i)}_{-i,\kappa,k} + \alpha_\kappa}{n^{(l_i)}_{-i,\kappa,\cdot} + K\alpha_\kappa} \cdot \frac{n^{(w_i)}_{-i,\kappa,k} + \beta}{n^{(\cdot)}_{-i,\kappa,k} + W\beta}, \kappa \in \{l, g\} \tag{2}$$

Consequently, the topic distribution of each location can be computed in Equation (3).

$$p(z_i = k | l_i) = \frac{n^{(l_i)}_{1,k} + \alpha_l}{n^{(l_i)}_{1,\cdot} + K\alpha_l} \tag{3}$$

4 Evaluation

4.1 Dataset

In this section, we experimentally evaluate the effectiveness of the proposed method. We report our experimental results on the following two real datasets.

The first come from Weibo [25] (all written in Chinese), which is a Chinese micro-blogging website. Akin to a hybrid of Twitter and Facebook, it is one of the most popular sites in China. Without loss of generality, we only focus on messages in Shanghai (the largest city in China and one of the largest cities in

the world by population) in 2015. Since most of the locations have been tagged for only a few times, we only selected those locations with a considerable number of posts within a pre-defined spatial range. The original data was preprocessed by filtering out stop words, then nouns and verbs were extracted as valid words with a Chinese POS(part-of-speech) tagger. Messages with less than three valid words were eliminated.

Our second dataset is from Yelp, which publish crowd-sourced reviews about local businesses, as well as the online reservation service and online food-delivery service. This dataset is publicly available [26], which is collected from Phoenix, which is a US city. In the Yelp dataset, each review has a location that is associated with a unique pair of latitude and longitude coordinates and a business name which is usually correspond to a restaurant, a hotel or an entertainment area. In order to share the same time span with the Weibo dataset, only reviews posted in 2015 is selected.

Some statistics about these two datasets are presented in Table 2.

Table 2. Dataset details

Description	**Weibo**	**Yelp**
Total number of Message	252173	661833
Total number of location	1088	43990
Average length of Message	110.32	33.30

4.2 Comparison Methods

We compare the proposed model LGLDA with the following other methods.

- **TF-IDF with K-means clustering (or TF-IDF).**
  In this method, Weibos are firstly preprocessed and presented as tf-idf weighted vectors and then aggregated by locations. Therefore, the feature vector of each location is summed by all documents tagged with that location. Finally, the feature vectors of each location are clustered by K-means. The center of each cluster represents a topic, and the weight of each element in the vector denotes the importance of the according keyword.

- **LDA model with location aggregation (or LDA).**
  In this method, topic and word distributions are firstly calculated by standard LDA algorithm with all documents, without considering the geo-tagged locations. After the global topics and word distribution in each topic are calculated, they are then aggregated by location. The topic distribution of each location is calculated as the average over all documents geo-tagged with that location.

– **Local LDA model (or LocalLDA).**
This method is similar to those in previous works. The topics are generated from locations instead of documents. If two words are from the same region, they are more likely to be clustered into the same topic. However, it is a simplified scenario, usually, if two words are close to each other in space, they are more likely to belong to the same region. In our dataset, the geographical distance of two words cannot be calculated due to the loss of the exact geo coordinations, but become a Boolean variable of whether they are tagged with the same location. Different from LGLDA model, in LocalLDA, there is only a local topic distribution, and all settings are kept the same with the LGLDA model.

4.3 Quantitative Measures

In order to make a comparison between different methods, several quantitative measures are used.

– **Perplexity**
Perplexity is used to evaluate the performance of topic modeling. Perplexity is the standard metric to evaluate the predictive power and generalizability of a topic model, and is monotonically decreasing with increasing likelihood of the test data set. Hence, a lower perplexity score indicates stronger predictive power.

$$perplexity(D) = exp\{-\frac{\sum_{d\in D} \log p(w_d)}{\sum_{d\in D} N_d}\} \tag{4}$$

where D is the test collection and N_d is the length of document d.

– **Location and Topic entropies**
The average topic entropy of each location and the average location entropy of each topic are used to measure the concentrativeness of discovered topics. Each location should have a compact distribution on topics, while each topic should concentrate on a small set of locations.

$$entropy_{topic} = \frac{1}{L}\sum_{L}\sum_{K} p_k^{(l)} \log p_k^{(l)} \tag{5}$$

$$entropy_{location} = \frac{1}{K}\sum_{K}\sum_{L} p_l^{(k)} \log p_l^{(k)} \tag{6}$$

where $p_k^{(l)}$ and $p_l^{(k)}$ are the estimated probabilities of topic k for location l and location l for topic k, respectively.

– **KL-divergence**
KL-divergence is used to measure the average distance of word distributions

of all pairs of topics. The larger the average KL-divergence is, the more distinct the topics are.

$$D_{KL}(p_i||p_j) = \sum_{L} p_i^{(k)} log \frac{p_i^{(k)}}{p_j^{(k)}} \tag{7}$$

where $p_i^{(k)}$ and $p_j^{(k)}$ are the estimated probabilities of topic k for location i and location j respectively.

4.4 Settings

In our experiments, the number of topics was set at K to 20 for all models (including K-means), α_l and α_g to 0.1, β to 0.1, γ_l and γ_g to 0.5 for all LDA-based models empirically and were run for 500 iterations.

For the LGLDA model, the local-global weight ratio was determined by model's perplexity as shown in Figure 2. As can be seen, in the Weibo dataset, with the gradually increasing value of λ from 0.1 to 20, the perplexity descended first and then ascended, and reached minimum at 0.6. Hence, 0.6 is the best value for λ for the Weibo dataset, and was used in our experiments. Performance on the other two metrics also confirmed this selection.

In the Yelp dataset, λ was determined with the same manner, and was finally set at 0.8. The optimal value of λ for the two dataset is quite close. In our future work, we would experiment against more different kinds of dataset to investigate whether the chosen value of λ is a coincidence or it is decided by the intrinsic properties of all user generated social media data.

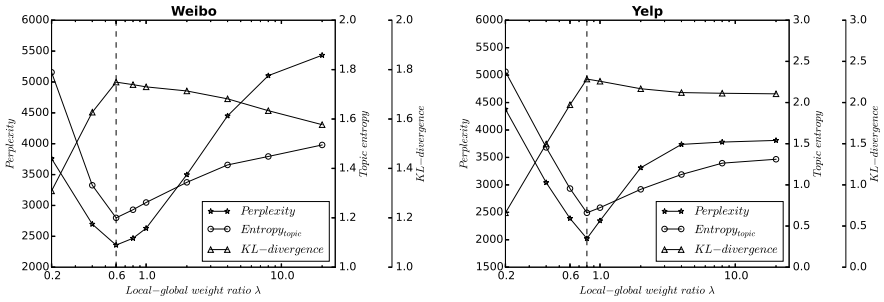

Fig. 2. The impact of the value of local-global weight ratio λ on model's performance.

4.5 Results

In this section, we experimentally evaluate the effectiveness of the LGLDA model, and compare it against the baseline methods.

Table 3. Examples of Weibo text with locality score sampled from location Waitan

Locality score	Weibo text
7.18	Cruising along Waitan while taking a close-up view of the Oriental Pearl Tower is a worthwhile trip.
0.91	Breakfast at Xitang, lunch at Hangzhou, dinner at Waitan. What an incredible day on the run.
0.02	To deal with a difficult customer in the afternoon, I'd have to get up early to do data analysis.

Locality score .

In order to validate the effectiveness of our LGLDA model to distinguish between local words and noise, we defined the locality score. The locality score is defined as the ratio of the average probability of words generated from the local and the global topic distribution, and it is calculated as in Equation (8).

$$Locality(d) = \frac{\sum_{w \in d} p(z_{e,w}, e_w = l)}{\sum_{w \in d} p(z_{e,w}, e_w = g)} \tag{8}$$

The locality score is a measurement of the relatedness of the messages to its respective local context or semantics. The higher the score is, the more representative the message is of its tagged location. In order to illustrate the usefulness of this measurement, we sorted all Weibos according to the locality score within each location, and three Weibos were sampled from the collection with distinguishable locality score with location tagged 'Waitan' and shown in Table 3.

As we can see, the result is quite in accordance with our expectation. Weibo with the highest score which includes several landmark names and location specific words is highly relevant, while the other two messages containing only few or no location featured words are weakly related or irrelevant. Therefore, our model has the ability to rank texts according to their location relevence.

Comparative results with baselines .

In this section, we use the quantitative measures described in Section 4.3 to evaluate the performances and show the superiority of our LGLDA model. The used quantitative measures include perplexity, KL-divergence and two kinds of entropies: topic entropy and location entropy.

Table 4 gives the comparative results of our LGLDA model with other baselines. As we can see, both in the Weibo dataset and in the Yelp dataset, our LGLDA model outperforms other baselines in almost all quantitative measures. Although Weibo and Yelp have distinctive business purposes and target users in different countries, while their datasets yield different statistic characteristics, our model is more preferable in both scenarios.

The LGLDA model achieves much lower perplexity for the reason that it can separate local and non-local words. Hence the words or the documents are better classified and organized by topics. With noisy words filtered out and only location

Table 4. Results of the comparative experiments

Dataset	Method	Perplexity	Topic entropy	Location entropy	KL-divergence
Weibo	LDA	6904.11	2.9680	59.5100	**2.2944**
	LocalLDA	5679.95	1.5156	31.2292	1.5694
	LGLDA	**2357.94**	**1.1998**	**24.2575**	**1.7494**
Yelp	LDA	6320.41	2.3669	44.9711	2.1769
	LocalLDA	4570.38	1.0965	20.8351	1.2330
	LGLDA	**2021.72**	**0.5608**	**10.6569**	**2.2892**

related words kept, the discovered topics are more distinct and representative of the local semantics, which is reflected by the topic and location entropies.

In Table 4, in the Weibo dataset, the KL-divergence of our model is lower than that of the original LDA model, and the reason may due to the location relevance constraint. Since in the LDA model, all messages are unseparated according to locations, the discovered topics can be formed by messages with different geo tags. Further more, because all messages are deemed as useful, the structure of the clusters discovered is different from the other LDA-based models.

Topic comparison of different methods .

In this section, we show the topics discovered by different methods with Weibo and Yelp datasets.

In all the models, we set the number of topics at 20, and since to list all of the keywords in each topic would have taken a lot of space while has little help to gain an useful insight, we only showed the top three topics discovered in each dataset here. The result is shown in Table 5.

Table 5. TOP3 topics discovered by Weibo and Yelp dataset

Weibo				Yelp			
TF-IDF	**LDA**	**LocalLDA**	**LGLDA**	**TF-IDF**	**LDA**	**LocalLDA**	**LGLDA**
work, company, mood, women, teacher	love, feeling, mate, inside, teacher	work, mood, love, children, phone	work, company, mood, phone, city	time, food, service, people, location	food, table, minutes, server, service	food, time, minutes, people, nice	food, service, restaurant, delicious, menu
university, teacher, effort, mood, paper	teacher, English, school, exam, culture	university, teacher, library, school, birthday	university, school, library, teacher, student	food, chicken, service, menu, love	steak, restaurant, dessert, bread, meal	steak, dinner, table, restaurant, server	steak, dessert, bread, cheese, salad
Waitan, city, night, Oriental Pearl Tower, restaurant	city, Waitan, Shanghai, Oriental Pearl Tower, international	Waitan, hotel, center, Oriental Pearl Tower, financial	Waitan, Oriental Pearl Tower, Chenghuang Temple, Nanjing Road, Huangpu River	hotel, stay, pool, desk, vegas	hotel, vegas, casino, desk, night	hotel, stay, time, vegas, night	hotel, vegas, casino, pool check

In the Weibo dataset, the first topic (2nd row) is composed of words with broader meanings, which can be viewed as noises, while the other two topics (3rd row and 4th row) contain the semantics of education and tourist attractions. As can be seen, keywords discovered by our LGLDA model achieve the best relevance while keywords by other methods include more or less noise (e.g. mood, city, etc.). To further drill down the result, in our LGLDA model, Topic1 together with Topic4 accounts for 98.8% of global topic distribution and contains a large amount of words, which is not related to any specific semantic locations. With these noisy words filtered out, locations can be covered by fewer topics. For example, for Oriental Pearl Tower, the weight of Topic3 in LGLDA is 0.933 compared with a mixed topic constitution discovered by LocalLDA (0.425 for Topic1, 0.414 for Topic3 and 0.161 for all others).

The result of the Yelp dataset is in accordance with the Weibo dataset. Firstly, the topics discovered by our LGLDA model are more distinct. The first topic and the second topic discovered is separable, since the first concentrates on the general aspect of meals or restaurants, while the second concentrates more on the detailed kinds of foods or dishes. However, the topics discovered by other models are more mixed up. Secondly, the keywords of topics discovered by our LGLDA model contain less noises. In contrast, keywords in the topics of other methods in more noisy (universal words may exist in different topics), therefore have impaired the semantic distinctness of the topics discovered.

5 Conclusion

This paper proposes a method, which combines local word filtering and geographical topic modeling into one framework. The proposed LDA-based model LGLDA can effectively distinguish between location related words and a variety of noisy daily interests by properly choosing the local-global weight ratio parameter in the Bayesian model. Results on Weibo collection show the effectiveness of our method over other baselines.

This initial work shows the potential for location-sensitive information retrieval and opens up several interesting future directions. Firstly, we would like to apply our models on other interesting data sources. For example, we can mine interesting geographical topics from the tweets associated with user locations in Twitter. Second, we would like to compare the topics discovered as local and global topics, and investigate the correlation between the topics discovered and human mobility pattern disclosed by other datasets such as cellular signaling and traffic sensor data.

References

1. Hofmann, T.: Probabilistic latent semantic indexing. *SIGIR*, pp. 50-57, ACM (1999).
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. *the Journal of machine Learning research*, vol. 3, pp. 993–1022 (2003).

3. Ashbrook, D., Starner, T.: Using GPS to learn significant locations and predict movement across multiple users. In *UbiComp*, pp. 275-286 (2003).
4. Mei, Q., Liu, C., Su, H., Zhai, C.: A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *WWW*, pp. 533–542, ACM (2006).
5. Wang, C., Wang, J., Xie, X., Ma, W.Y.: Mining geographic knowledge using location aware topic model. In *GIR*, pp. 65–70, ACM (2007).
6. Backstrom, L., Kleinberg, J., Kumar, R., Novak, J.: Spatial variation in search engine queries. In *WWW*, pp. 357–366, ACM (2008).
7. Palma, A.T., Bogorny, V., Kuijpers. B., Alvares, L.O.: A clustering-based approach for discovering interesting places in trajectories. In *SAC* (2008).
8. Li, H., Li, Z., Lee, W.C., Lee, D.L.: A probabilistic topic-based ranking framework for location-sensitive domain information retrieval. In *SIGIR*, pp. 331–338, ACM (2009).
9. Sizov, S.: Geofolk: latent spatial semantics in web 2.0 social media. In *WSDM*, pp. 281–290, ACM (2010).
10. Mathioudakis, M., Koudas, N.: Identifying, attributing and describing spatial bursts. In *Proceedings of the Vldb Endowment*, pp. 1091–1102, ACM (2010).
11. Eisenstein, J., Connor, B.O., Smith, N.A., Xing, E.P.: A latent variable model for geographic lexical variation. In *EMNLP*, pp. 1277–1287, ACM (2010).
12. Cheng, Z., Caverlee, J., Lee, K.: You are where you tweet: a content-based approach to geo-locating twitter users. In *CIKM*, pp. 759–768, ACM (2010).
13. Li, Z., Ding, B., Han, J., Kays, R., Nye, P.: Mining periodic behaviors for moving objects. In *SIGKDD*, pp. 1099–1108, ACM (2010).
14. Yin, Z., Cao, L., Han, J., Zhai, C., Huang, T.: Geographical topic discovery and comparison. In *WWW*, pp. 247–256, ACM (2011).
15. Ye, M., Yin, P., Lee, W.C., Lee, D.L.: Exploiting geographical influence for collaborative point-of-interest recommendation. In *SIGIR*, pp. 325–334, ACM (2011).
16. Hong, L., Ahmed, A., Gurumurthy, S., Smola, A.J., Tsioutsiouliklis, K.: Discovering geographical topics in the twitter stream. In *WWW*, pp. 769–778, ACM (2012).
17. Bauer, S., Noulas, A., Seaghdha, D.O., Clark, S., Mascolo, C.: Talking places: Modelling and analyzing linguistic content in foursquare. In *SocialCom/PASSAT*, pp. 348–357, IEEE (2012).
18. Hu, B., Ester, M.: Spatial topic modeling in online social media for location recommendation. In *RecSys*, pp. 25–32, ACM (2013).
19. Hu, B., Jamali, M., Ester, M.: Spatio-temporal topic modeling in mobile social media for location recommendation. In *ICDM*, pp. 1073–1078, ACM (2013).
20. Ahmed, A., Hong, L., Smola, A.J.: Hierarchical geographical modeling of user locations from social media posts. In *WWW*, pp. 25–36, ACM (2013).
21. Yuan, Q., Cong, G., Ma, Z., Sun, A., Thalmann, N.M.: Who, where, when and what: discover spatio-temporal topics for twitter users. In *SIGKDD*, pp. 605–613, ACM (2013).
22. Kim, Y., Han, H., Yuan, C.: TOPTRAC:Topical Trajectory Pattern Mining. In *SIGKDD*, pp. 587–596, ACM (2015).
23. Liu, Y., Ester, M., Hu, B., Cheung, D.W.: Spatio-temporal topic models for check-in data. In *ICDM*, pp. 889–894, IEEE (2015).
24. Wu, F., Li, Z., Lee, W.C., Wang, H., Huang, Z.: Semantic annotaion of mobility data using social media. In *WWW*, pp. 1253–1263, ACM (2015).
25. https://en.wikipedia.org/wiki/Sina_Weibo.
26. https://www.yelp.com/dataset_challenge.