Location-Based Topic Evolution

Haiqin Yang[1], **Shouyuan Chen**[1], **Michael R. Lyu**[1], **Irwin King**[1,2]
[1]Department of Computer Sciences and Engineering
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong
{hqyang, sychen, lyu, king}@cse.cuhk.edu.hk
[2]AT&T Labs Research
201 Mission Street
San Francisco, CA
irwin@research.att.com

ABSTRACT
As the advance of mobile technologies, geographical records can be easily embedded in the data to form the *location-associated documents*. For example, in Twitter, the location of tweets can be identified by the GPS locations or IP addresses from smart phones. In Flickr, photos may be tagged and recorded with GPS locations. With the geographical information, it is more likely to model users' interests in different regions so as to determine the corresponding marketing strategy. Due to its potential in providing personalized and context-aware services, several pieces of work have started to explore in this area. One stream of work tries to discover users' interest topics from location-associated documents. These models work under the assumption that words close in geographical positions are likely to be clustered into the same geographical topic. However, they attain this in a static mode. That is, they do not consider the evolution of the topics. In addition, they have to specify the total number of topics for the corpus in advance. In order to utilize the geographical information and to model the change of topics, we propose a location-based topic evolution (LBTE) model to tackle the above issues. Main advantages of our model lie that it can reveal the appearance and disappearance of the topics in different regions. Moreover, topics can be automatically determined based on the location-associated documents and its total number is not restricted to a preset value. Finally, we conduct a series of experiments on both synthetic and real-world datasets to demonstrate the merits of our proposed LBTE model in capturing users' interest topics.

Author Keywords
Location-based service, topic evolution, functional space

ACM Classification Keywords
H.4 Information Systems Applications: Miscellaneous; I.2 Artificial Intelligence: Learning

MLBS'11, September 18, 2011, Beijing, China.

General Terms
Algorithms, Experimentation, Verification.

INTRODUCTION
Due to the advance of mobile technologies, such as smart phone, GPS, etc., location information can be easily incorporated into the corresponding data [4, 12, 28]. The location information may be a geographical record, which represents a unique location on the Earth. Currently, these kinds of applications become prevalent. For example,

- In Twitter [1], tweets can be posted with the GPS records or IP addresses which also identify the user's current position. These geographical records with text content have been utilized to detect real-time events, such as estimating Typhoon trajectory or Earthquake location [20].
- In Flickr [2], over 100 million photos are tagged and included explicitly with their GPS locations. The geo-tagged information can provide users' common interests, culture, etc., in the corresponding regions. For example, as illustrated in Figure 1, photos with geo-tagged information can identify restaurants or earthquake and tsunami at that position.

Overall, the geographical records can be embedded in various documents, such as user message, user posts, tags, to construct the *location-associated documents*. Usually, common documents are generated under some topics, where each topic is characterized by a distribution over words [25]. With geographical information, the relationship between documents can be incorporated. This can help to model the topics among documents more accurately. In addition, these documents may implicitly reveal users' interests through the learned topics. With geographical information, we can identify users' interests in the corresponding regions and then determine appropriate marketing strategy for them [6].

Due to the promise of analyzing location-associated documents, several pieces of work have been investigated to model geographical topics [24, 25] or geo-tagged photos [5, 21]. The essential of these models assumes that words close

[1]http://twitter.com/
[2]http://www.flickr.com/

Figure 1. Examples of photos with geo-tagged information in Flickr.

in geographical positions are likely to be clustered into the same geographical topic. Although they work well in finding regions of interests, they are still limited in the following issues:

- *Topics Generation Modeling*. Each topic must be generated and withered away on a position at a specific period. This can be modeled to know more about the scenario of discussing the topics. However, previously proposed methods lack the ability in modeling the appearance and disappearance of topics.

- *Topics Evolution Modeling*. Topics are not fixed all the time. For example, a tweet may initialize the topic, "earthquake", at a certain position. This topic may be further discussed at the same region and spread to other regions until it was withered away after a period of time. With the help of location-associated documents, we can model the changes of users' interests more accurately. However, none of the previous work can tackle this issue and we aim to develop an effective model to capture the change among topics.

- *Topics Number Determination*. Previously proposed models preset the total number of topics before the learning procedure. However, determining the number of topics needs prior knowledge. In order to make it automatically, one may consider adopting nonparametric methods, e.g., Hierarchical Dirichlet processes [23], to solve it. When dealing with the location-associated documents, none of the previous work consider this issue and we decide to design an efficient way to automatically model the topics without restricting its number being smaller than a preset value.

To tackle the above issues, in this paper, we propose a location-based topic evolution (LBTE) model to capture the change of users' interest topics within different regions or time periods. More specifically, the LBTE models users' interests through topic evolution which is innerly mastered by a collection of unknown, but countably infinite continuous functions. These functions capture users' interest topics in two situations:

1) The change of topics varies smoothly with the variant of regions;

2) The change of topics varies smoothly with the variant of regions at different periods.

More importantly, through functional domains definition, our LBTE model can allow for the appearance and disappearance of topics. Moreover, topics can be determined automatically without predefining a maximum value for restricting the total number of topics before the training.

RELATED WORK

In this section, we address work related to topic modeling with analysis on spatially distributed data such as GPS positions, demographics information, etc.

Location-based Analysis

Recently, researches on location-based service (LBS) have been conducted due to a wide range of potential applications, such as personalized marketing strategy analysis, personalized behavior study, context-aware analysis, etc. These methods try to capture users' patterns through various algorithms from their sequentially behaviors [9, 13, 27, 28].

Typically, in [9], several heuristic methods have been proposed to discover users' trajectory patterns to concisely describe users' frequent behaviors in both space and time. In [27], the association rules related algorithms have been proposed the to mine mobile sequential pattern by considering moving paths and adding the moving path between the left hand and the right hand in the content of rules. In [13], a cluster-based temporal mobile sequential pattern mine algorithm is proposed to discover users' temporal mobile sequential patterns. In [19], tags and GPS metadata in the images from Flickr are analyzed to extract place and event semantics by the method of scale-structure identification.

Topic Modeling

Topic modeling is a classical problem in information retrieval and text mining. It usually models topics through a word distribution. Typical methods include probabilistic latent semantic analysis (pLSA) [10] and latent Dirichlet allocation (LDA) [3]. Various extensions have been conducted to apply in analyzing the spatially distributed data [8, 14, 22, 26]. In the following, we emphasize several work close to the idea in this paper.

In [8], a probabilistic topic model based on LDA are proposed to discover individuals' daily routines from human locations. The proposed method is verified on the Reality Mining dataset [7] from mobile phone users. In [14], a probabilistic approach is proposed to model the spatiotemporal theme patterns in weblogs. In [22], a framework, called GeoFolk, is proposed to combine both text and spatial information to construct better algorithms for content management,

Table 1. Notations

	Description
V	Vocabulary includes all words
$\mathcal{C} = \{(\mathbf{w}_l, G_l(G_l^{t_l}))\}_{l=1}^N$	A corpus of documents consist of N location-associated documents
l	A document l consists of text and location, i.e., $(\mathbf{w}_l, G_l(G_l^{t_l}))$
$\mathbf{w}_l$	Text in document l, $\mathbf{w}_l = (w_{l1}, w_{l2}, \ldots)$
G_l $(G_l^t) \in X$	Location of document l (at time t)
X	The domain of underlying hidden functions defines on the region (time).
Y	The range of underlying hidden function values. A function value corresponds to a topic.
$h : X \to Y$	Underlying hidden functions model the change of topics.
$[K]$	An integer set consists of $1, 2, \ldots, K$

retrieval, and sharing in social media. In [26], a joint model, the latent geographical topic analysis (LGTA), is proposed to combine both location and text information. Both GeoFolk and LGTA are evaluated on the public data from Flickr.

In summary, previously proposed methods have taken context-aware information, such as time and location, into account in the corresponding applications to reveal users' patterns. When including text information, topic models are more suitable to seek users' interest topics. However, these methods do not explicitly consider modeling the appearance and disappearance of a topic. Moreover, they have to specify the number of topics beforehand. These insufficiency motivates our work in this paper.

PROBLEM SETUP

In this section, we define the problem of location-based topic evolution. To make succinct, the notations used in the paper are depicted in Table 1.

Suppose there are N location-associated documents in a corpus $\mathcal{C}$, each location-associated document is encoded with its location or with time simultaneously. Hence, we can obtain $\mathcal{C} = \{(\mathbf{w}_l, G_l)\}_{l=1}^N$ or $\mathcal{C} = \{(\mathbf{w}_l, G_l^{t_l})\}_{l=1}^N$, where each document l consists of a set of words $\mathbf{w}_l$, the words are from vocabulary V, and is embedded with its location information G_l, or location-time information, $G_l^{t_l}$. In the following, we use G_l to denote either of the case.

The problem of *location-based topic evolution* is defined as follows. Given a corpus of location-associated documents, $\mathcal{C}$, we are interested in modeling *the topics of data with an unknown number of topics and parameters*.

MODEL–LOCATION-BASED TOPIC EVOLUTION

In this section, we propose a novel location-based topic evolution model, namely LBTE, to combine the topic change with the region (time) change in a uniform framework.

Assumptions

Popular topic models, such as pLSA [11] and LDA [3], model documents using mixture of topics and represent them by a low dimensional representation. By assuming documents are abstracted by some latent semantic topics, these models have been extended to develop the topic structure by including time or location information [2, 25]. In developing geographical topics, the assumptions are further extended to meet the requirement of the location-based applications: If two words are close to each other in space, they are more likely to belong to the same region. If two words are from the same region, they are more likely to be clustered into the same topic [2, 25].

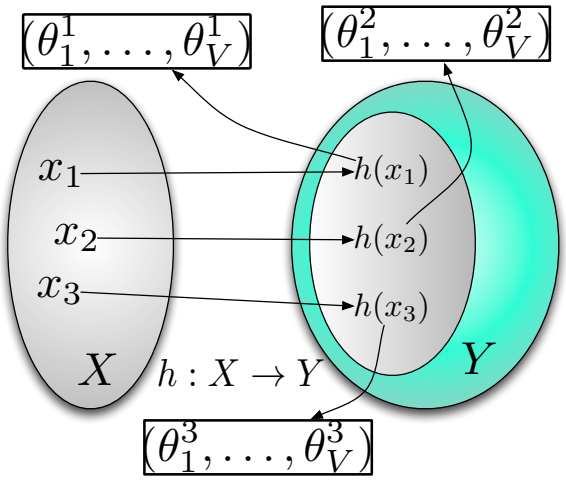
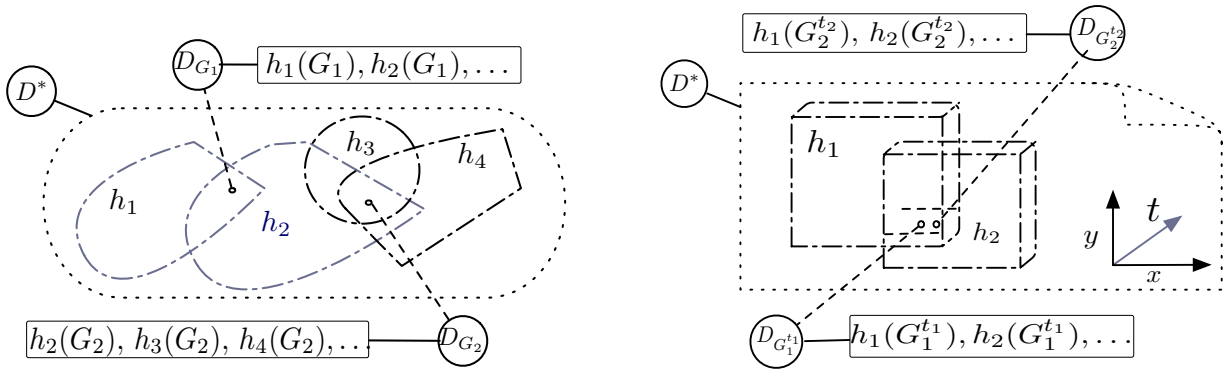


Figure 3. Illustration of a underlying hidden function.

Differently, we do not directly pose the above assumptions. On the contrary, we assume the location-associated documents are generated from a collection of countably infinite continuous underlying hidden functions, where each underlying hidden function is $h : X \to Y$. This assumption is close to that in [18], which only focuses on the theoretical study of the model. Here, we emphasize the assumption of the model for the location-based applications.

As illustrated in Figure 3, we assume the location-associated documents are generated from unknown number of topics, where each topic is represented by $(\theta_1, \ldots, \theta_V)$. It should be noted that each topic is determined by the value of a underlying hidden function. Hence, it can represent the change of topic distribution through underlying hidden functions. Moreover, any one of the underlying hidden functions, h, is defined on X, which represents the life-span of a topic, e.g., regions or regions over a period. Furthermore, the collection of functions is constructed from a probability measure D^*,



(a) Evolving with Regions

(b) Evolving with Regions and Time

Figure 2. A representation of topics evolving with regions and time.

which is generated from from Dirichlet process parameterized by μ^* in a finite base measure over Y:

$$D^* \sim \text{DP}(\mu^*) \tag{1}$$

Model

Now we focus on how to model the generation of topics based on the assumption of the underlying hidden function collection generation scheme from D^*. Figure 2 illustrates two cases of the underlying hidden functions to generate the topics. It is noted that the domains of the functions can be regions (see Fig. 2(a)) or regions at different periods (see Fig. 2(b)), respectively. More specifically,

- In Figure 2(a), D_{G_1} illustrates a Dirichlet process on location G_1, where the domains of functions such as h_1 and h_2 include the location G_1. Meanwhile, the values of $h_1(G_1)$ and $h_2(G_1)$ determine the distribution of words in topic1 and topic2, respectively. Correspondingly, the topics at the location G_2 are determined by the values of all functions of $h_2(G_2)$, $h_3(G_2)$, and $h_4(G_2)$, whose domains include the location G_2. The corresponding documents can be generated from D_{G_2}, also a Dirichlet process. It should be emphasized again that the domain of h_i, $i = 1, 2, 3, 4$ corresponds to the life-span of four specific topics, respectively. For example, since the domains of h_3 and h_4 do not include the location of G_1, but include the location of G_2, the corresponding topics do not appear in G_1, but appear in G_2.
- Similarly, in Figure 2(b), $D_{G_1^{t_1}}$ and $D_{G_2^{t_2}}$ model Dirichlet processes on location G_1 at time t_1 and on location G_2 at time t_2, respectively. This figure is easier to illustrate the appearance and disappearance of a topic. For example, the beginning and end of the function domain, h_1, correspond to the "appearance" and "disappearance" of the topic at time axis.

In summary, the topics of models are determined by the function values. All the topics are generated from a Dirichlet process D^*. Restricting and renormalizing D^* to include functional atoms whose domain containing the location, such as G_1 and G_2, corresponds to projecting each function at the location. It should be noted that a good property of the above modeling is that the marginal distribution of D_G at a location G is still a Dirichlet process,

$$D_G \sim \text{DP}(\mu_G), \tag{2}$$

where μ_G is a measure defined on Ω by

$$\mu_G(F) = \mu^* \left(\{h : G \in X,\ h(G) \in Y,\ Y \in \Sigma\}\right).$$

In addition, each topic is characterized by the value of function, such as $h(G_1)$ and $h(G_2)$. Based on the distribution of words on a topic, we can then generate the corresponding documents. Hence, the generative process of observation $\{(\mathbf{w}_l, G_l)\}_{l=1}^N$ is

$$h_l | D_{G_l} \sim D_{G_l}, \tag{3}$$

$$\theta_l | h_l = h_l(G_l), \tag{4}$$

$$\mathbf{w}_l \sim H_{\theta_l}(\cdot), \tag{5}$$

where in the above, $1 \le l \le N$ and H_θ is a probability distribution parameterized by θ. The document $\mathbf{w}_l$ at location G_l is drawn from a mixture component with parameter θ_l whose value is determined by the underlying hidden function h_l.

Inference

To infer the model, we first determine the functional assignment of domains, which corresponds to the life-span of the topic, by selecting an appropriate prior distribution. Following Kolmogrov extension theorem [17], here, we define the distribution of function domains as the probability that the region is defined in the function domain:

$$g\left(\{G_l\}_{l=1}^N\right) = \Pr\left(\{G_l\}_{l=1}^N \subseteq X\right).$$

More specifically, it is defined by

$$g\left(\{G_l\}_{l=1}^N\right) = \exp\left(-\tau \max_{j,k}\{c(G_j, G_k)\}\right), \tag{6}$$

where $\tau > 0$ controls the magnitude of the probability, the function, $c(G_j, G_k)$, defines the closeness of two regions G_j and G_k (or with time).

Next, we develop a Gibbs sampler to perform the approximate inference from observed documents. We assume $h^*_{1:K}$ be the unique functions among $h_{1:N}$ of (3) and each unique function h^*_l appears n_l times. Hence, we have

$$K \leq N, \quad \text{and} \quad \sum_{i=1}^{K} n_i = N.$$

We denote the assignments of functions as d_i, where $1 \leq d_i \leq K$. Hence, we have

$$h^*_{d_i} = h_i.$$

The predictive distribution is defined as

$$d, \theta \mid \{d_{1:(N-1)}, \theta_{1:(N-1)}\} \tag{7}$$

We now aim at deriving it for location-associated documents with text $\mathbf{w}_l$ on location G_l (at time t_l). Due to the DP assumption on D^*, the posterior of D^* given $\{z_{1:(N-1)}, \theta_{1:(N-1)}\}$ is a *mixture of Dirichlet process* (MDP) [1]. Through simplifying and marginalizing out DPs, we can derive the predictive distributions as follows:

$$d \mid \{d_{1:(N-1)}, \theta_{1:(N-1)}\} = \mathop{\mathbb{E}}_{b^G_{1:K}, \theta^G_{1:K}} A \tag{8}$$

$$\theta \mid \{d_{1:(N-1)}, \theta_{1:(N-1)}\} = Z^G_d(\cdot), \tag{9}$$

where

$$A = \left[\frac{\mu_G(V)\delta_{K+1} + \sum_{i=1}^{K} b^G_i n_i \delta_i}{\mu_G(V) + \sum_{i=1}^{K} b^G_i n_i}\right] \tag{10}$$

$$b_i \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p^G_i), \tag{11}$$

$$\theta_i \overset{\text{i.i.d.}}{\sim} Z^G_i(\cdot) \tag{12}$$

and δ_{K+1} indicates the event of assigning a new function. The values p^G_i and $Z^G_i(\cdot)$ are induced from G^* by

$$p^G_i = \Pr(G \in X | \{G_{1:(N-1)}\} \subseteq X) = \frac{g(\{G_{1:(N-1)}, G\})}{g(\{G_{1:(N-1)}\})}, \tag{13}$$

$$Z^G_i(\theta) = \Pr(h(G) = \theta | h(G_1) = \theta_1, \ldots, h(G_{N-1}) = \theta_{N-1}). \tag{14}$$

Hence, by treating $b^G_{1:K}$ and $\theta^G_{1:K}$ as auxiliary variables, we make it tractable sample d from the predictive distribution. Actually, the auxiliary variables $b^G_{1:K}$ and $\theta^G_{1:K}$ contain neat interpretation: $b^G_{1:K}$ is a random event that the domain of a function h includes G given that its domain include $G_{1:(N-1)}$; while $\theta^G_{1:K}$ is a random variable which is equal to $h(G)$ conditioned on $h(G_i) = \theta_i$.

Now, we develop the Gibbs sampler to perform an approximate inference. It should be noted that the Gibbs sampler only maintains two kinds of information: all assignments, d_i, and the number of occurrence of different functions, n_i, where $1 \leq i \leq K$, K is the number of represented functions and can be changed in the iterations. Hence, the sampling procedure is as follows:

For $k \in [K]$

1. Resample b^G_k as defined in (11) to determine whether the domain of h^*_k contains G_N;
2. Draw model parameters θ^G_k as defined in (12).

It is noted that the probability of assigning the N-th sample to existing functions is

$$\Pr(d_N = k) \propto b^G_k n_k H_{\theta^G_k}(\mathbf{w}_N), \tag{15}$$

while the probability of assigning the N-th sample to a new function is

$$\Pr(d_N = K + 1) \propto \mu_{G_N}(V) H(\mathbf{w}_N), \tag{16}$$

where $F(\mathbf{w}_N) \triangleq \mathbb{E}_{\theta \sim M_{G_N}(\cdot)} H_\theta(\mathbf{w}_N)$ and $M_{G_N}(\cdot)$ is the marginal distribution of base process at G_N.

EXPERIMENTS

In this section, we conduct experiments on both synthetic and real-world dataset to demonstrate the merits of our proposed model.

Synthetic Dataset

We first present the generation procedure and the results of synthetic dataset. The generation of the synthetic dataset consists of two steps:

- **Topics Generation:** It includes the initialization of topics and the scheme of topics evolution.
  1. **Topics Initialization:** Two topics are generated at the beginning. Each topic contains a center which follows a Gaussian distribution with zero mean and variance 20, i.e., $\mathcal{N}(0, 20)$. The parameter of each topic follow a Gaussian distribution of $\mathcal{N}(0, 10)$.
  2. **Topics Evolution:** At each time stamp, the old topics die off with the probability of 40%, while new topics occur following the Poisson distribution with parameter 0.8. It should be noted that the expectation of the number of topic is 2 at each time stamp.
- **Location-associated Documents Generation:** At each time stamp, we generate 10 documents for each topic. The location of each document follows the uniform distribution at the center of the topic with the radius of 5.

  The values of the documents are generated following a Gaussian distribution $\mathcal{N}(\nu, 1)$, where ν is the topic parameter.

In the experiment, we generate totally 730 location-associated documents at 30 time stamps, consisting of 28 different topics. In our LBTE, we have to define the function c in (6) to determine the topic assignment. Here, for simplicity, we define it as $c(G^{t_a}_a, G^{t_b}_b) = \|G_a - G_a\|_2 + |t_a - t_b|$. Hence, two documents in closer locations or in short time stamps are more likely to be generated from the same topic.

Since we have the ground truth for the synthetic dataset, we adopt the criterion of variation of information [15] to evaluate the model performance. The criterion of variation of

Data Set	# images	# unique tags	# total tags
Landscape	1505	243	2313
Activities	11868	232	2381
National Park	2109	257	2374

Table 2. The statistics of the datasets.

information is defined as follows:

$$d_{VI}(\mathcal{C}_1, \mathcal{C}_2) = H(\mathcal{C}_1) + H(\mathcal{C}_2) - 2I(\mathcal{C}_1, \mathcal{C}_2), \qquad (17)$$

where H and I denote the entropies of and the mutual information between the two clusters, respectively. d_{VI} measures the distance between two clusterings in terms of the information difference between the two clusters. Hence, the lower the value of d_{VI}, the better the performance is.

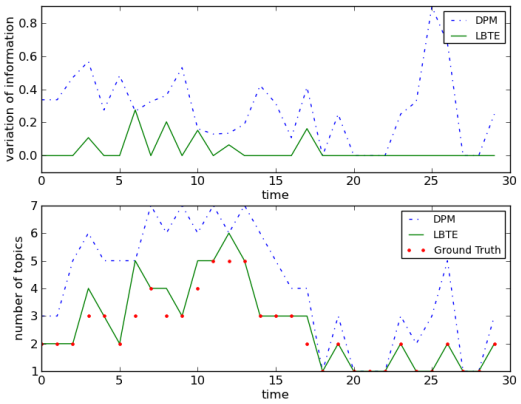


Figure 4. Results on synthetic data. The lower the variation of information, the better the model is.

The experiment is evaluated on the synthetic dataset by our LBTE and comparing with the Dirichlet process mixture (DPM) model [16] as a benchmark method. Figure 4 shows the variation of information and the number of topics generated from our LBTE and the DPM model. It is shown that our LBTE outperforms the DPM at all the time stamps. Especially, LBTE recovers the true topics and achieves zero variation of information when the time stamp is greater than 18. On the contrary, the DPM cannot capture the change of topics and still fluctuates with time.

Real-World Dataset

In this section, we evaluate the LBTE on Flickr dataset. The images with GPS locations are crawled through Flickr API [3], where the Flickr API supports search criteria of tag, time, GPS range, etc.. We select three representative datasets, including Landscape, Activity, and National Park, which are more related to travel, in the evaluation. The photos are crawled in the time span from 2009/01/01 to 2010/01/01 and only kept in USA territory. We remove tags occurring less than 15 times in the datasets. The statistics of the datasets are listed in Table 2. More details about the datasets are explained in the following:

[3] http://www.flickr.com/services/api

- For Landscape dataset, we crawl the images containing keyword *landscape* around USA.
- For Activities dataset, we crawl the images containing keywords *surfing* and *hiking* around USA.
- For National Park dataset, we crawl the images containing keyword *nationalpark* around USA.

We compare the following methods in the experiment:

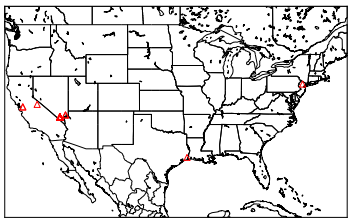
- DPM: Dirichlet Process Mixture [16]. This method is adopted as a benchmark method. It automatically learns the topics only from tags information.
- LBTE: Our proposed Location-Based Topic Evolution method. We set the parameter τ in (6) to 0.5 and let Z_i^G in (14) be a symmetrical dirichlet distribution parameterized by β, i.e., $Z_i^G(\cdot) = \text{Dir}(\beta)$, where $\beta = 0.1$. In addition, we let $\mu_G(V) = 1.0$ and the number of iteration in Gibbs sampling be 1000.

Since the discovered topics by different methods may be different, we seek similar topics between LBTE and DPM by calculating the cosine similarity of two topics by different methods. The highest cosine similarity value of two topics is set as the same topic learned by two models.

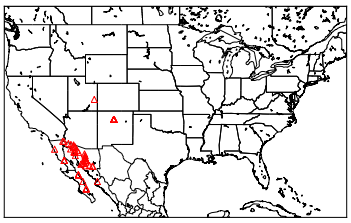
Topic Discovery from Landscape Dataset

In landscape dataset, our LBTE model can automatically learn nine topics. On the contrary, the DPM only attains six topics. We list ten representative tags with their weights from three representative topics discovered by DMP and four representative topics discovered by LBTE in Table 3 and show the corresponding geo-tagged photos in Figure 5. From these results, we have the following observations:

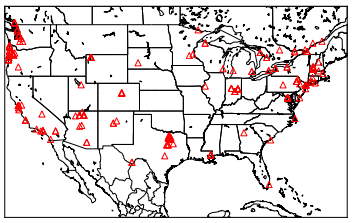
- From Table 3, it is noted that DPM does not consider the location information, which yields including photos in Oregon and Yellow Stone into the same topic. Differently, the LBTE can separate these two topics clearly.
- By examining the details in Table 3, we can see that the discovered topic 1 and topic 2 by DPM and by LBTE are nearly the same. They also share similar tags, such as landscape, desert, mexico, etc. The slight difference lies in the weights for the corresponding tags. For the representative topic 3 discovered by DPM, it contains tags which also appear in the topic 3 and topic 4 discovered by LBTE. These tags include landscape, nature, sky, travel, etc., which are more related to travel.
- The above observations can be clearly shown in Figure 5. Our LBTE utilizes the location information sufficiently and discovers the topics based on the regions of the photos, where the results are more explainable. Contrarily, the DPM seeks the topics without location information. Especially, as in illustrated in Fig. 5(c), the discovered topics by the DPM is scattered and lack of interpretation.



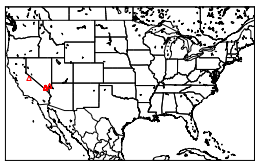
(a) Topic 1 (DPM)



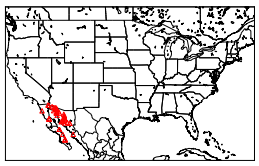
(b) Topic 2 (DPM)



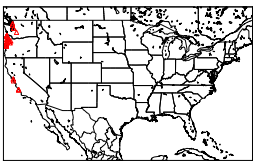
(c) Topic 3 (DPM)



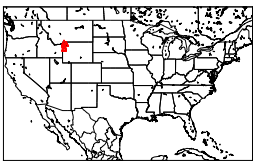
(d) Topic 1 (LBTE)



(e) Topic 2 (LBTE)

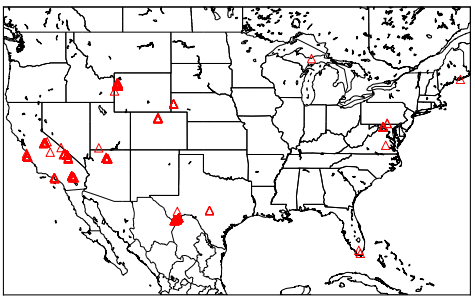


(f) Topic 3 (LBTE)

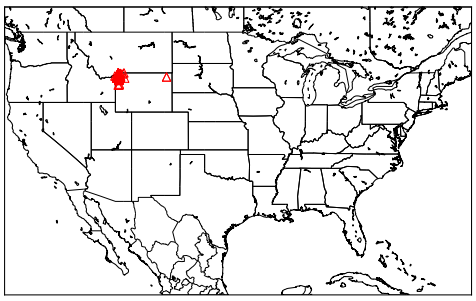


(g) Topic 4 (LBTE)

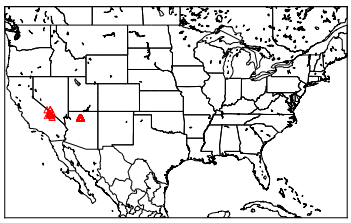
Figure 5. Photos in the topics discovered by DPM and LBTE on landscape dataset.



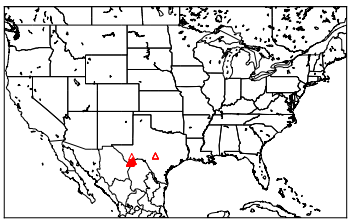
(a) Topic 1 (DPM)



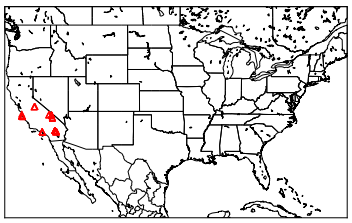
(b) Topic 1 (LBTE)



(c) Topic 2 (LBTE)



(d) Topic 3 (LBTE)



(e) Topic 4 (LBTE)

Figure 6. Photos in the topics discovered by DPM and LBTE on national park dataset.

DPM			LBTE			
Topic 1 Nevada Desert	Topic 2 Mexico	Topic 3	Topic 1 Nevada Desert	Topic 2 Mexico	Topic 3 Oregon	Topic 4 Yellow Stone
landscape 0.026810	landscape 0.083929	landscape 0.141658	landscape 0.026492	landscape 0.082290	landscape 0.142061	landscape 0.160494
desert 0.026475	paisaje 0.082143	nature 0.048618	desert 0.026492	paisaje 0.082290	coast 0.108635	wyoming 0.131687
timeofday 0.026475	mexico 0.082143	usa 0.026583	timeofday 0.026492	mexico 0.082290	oregon 0.086351	yellowstonenationalpark 0.102881
desertlandscape 0.026475	landschaft 0.057143	california 0.022036	desertlandscape 0.026492	landschaft 0.057245	ocean 0.069638	nature 0.069959
lasvegas 0.026475	tanawin 0.057143	sunset 0.021686	lasvegas 0.026492	tanawin 0.057245	beach 0.066852	geotagged 0.057613
nevada 0.026475	paisagem 0.057143	sky 0.021686	nevada 0.026492	paisagem 0.057245	northwest 0.058496	travel 0.053498
nevadausa 0.026475	landschap 0.057143	travel 0.021336	nevadausa 0.026492	landschap 0.057245	pacific 0.052925	usa 0.037037
nevadadesert 0.026475	landskap 0.057143	water 0.020637	nevadadesert 0.026492	landskap 0.057245	coastline 0.047354	national 0.037037
nevadastatepark 0.026475	Mekcnka 0.057143	photo 0.018888	nevadastatepark 0.026492	Mekcnka 0.057245	sunset 0.044568	sunset 0.032922
redrock 0.026475	méxico 0.057143	canon 0.015740	redrock 0.026492	méxico 0.057245	sky 0.030641	trip 0.032922

Table 3. Topics discovered in landscape dataset from DMP and LBTE.

Topic Discovery from National Park Dataset

In the national park dataset, our LBTE has discovered twelve topics; while the DPM only discovers five topics. Here, we list ten representative tags with their weights from one representative topic discovered by DPM and four representative topics discovered by LBTE in Table 4 and show the corresponding geo-photos of the topics in Figure 6. We have the following observations:

- From Table 4, it clearly shows that LBTE finds the topics based on four different national parks. For example, the topic 1 discovered by LBTE is related to yellow stone since tags such as yellowstone, shouthdakota, appear in this topic. Similarly, in the topic 2 discovered by LBTE, tags such as grandcanyon, grandcanyonnationalpark can represent the content of this topic clearly. For the topic 3 and topic 4, tags such as bigbend and joshuatree are the representative keywords. On the contrary, DPM mixes these words together and cannot distinguish them well.

- Results in Figure 6 again clearly show that topics discovered by LBTE are around the four national parks; while DPM includes nearly all four parks in the same topic.

Topic Discovery from Activities Dataset

In the activities dataset, DPM discovers nine topics; while our LBTE discovers fourteen topics. We list ten representative tags with their weights from three representative learned from DPM and four representative topics learned from LBTE in Table 5. It is interesting to know that LBTE separates "Hiking" into two topics according to the locations; while DPM does not consider this issue and combine them together.

By examining the details in Table 5, we can notice that the topic 2 discovered by LBTE is related to some hiking tracks closer to wyoming and California; while the topic 4 is related to some hiking tracks close in Arizona or desert areas. At the same time, the topic 2 discovered by DPM finds all places related to hiking in the whole country of USA.

In summary, our LBTE usually generates more topics than DPM. This is reasonable since our LBTE discover the topics by considering the location information on the tags. This requires a slight constriction on the topic generation and separates more topics. Typical examples can be viewed in the topic 3 discovered by DPM and the topic 3, 4 discovered by LBTE, and the result from national park and activities datasets. However, we find that increasing the number of total topics is worthy since we can interpret the generated topics clearer. This has been evidenced in the experimental results.

CONCLUSION

In this paper, we develop a location-based topic evolution (LBTE) model to capture the appearance and disappearance of geographical topics from the location-associated documents. The advantages of the LBTE include 1) automatically modeling the number of total topics; 2) automatically modeling the appearance and disappearance of topics; 3) inferencing the model by Gibbs sampling, which is simple and succinct. The experimental results on synthetic and real-world datasets demonstrates our proposed LBTE utilizes the location information sufficiently in discover the topics and the topic evolution.

We hope this work can inspire several directions on location-

DPM	LBTE			
Topic 1	Topic 1 Yellow Stone	Topic 2 Grand Canyon	Topic 3 Big Bend	Topic 4 Joshua Tree
nationalpark 0.130981	nationalpark 0.335780	nationalpark 0.166172	nationalpark 0.064000	nationalpark 0.068482
nature 0.095983	yellowstone 0.297248	desert 0.069733	scenery 0.029647	joshuatree 0.066832
park 0.094211	wyoming 0.282569	usa 0.066024	waterfall 0.025882	california 0.066832
yellowstone 0.082841	southdakota 0.014679	grandcanyon 0.052671	bigbendnationalpark 0.024000	beach 0.057756
america 0.058771	wildlife 0.014679	deathvalley 0.052671	bigbend 0.024000	keyesranch 0.056931
yellowstonenationalpark 0.058771	2009 0.012844	landscape 0.048220	westtexas 0.024000	photos 0.056931
americasfirst 0.058328	elk 0.007339	arizona 0.046736	texas 0.024000	bidsur 0.056931
wyoming 0.022445	spring 0.007339	grandcanyonnationalpark 0.023739	canyon 0.021176	venice 0.056931
yosemite 0.021412	vacation 0.005505	us 0.023739	summer 0.021176	hermosa 0.056931
osprey 0.021116	canyon 0.003670	geology 0.022997	sunset 0.018353	noahpurifoy 0.056931

Table 4. Topics discovered in National Park dataset from DMP and LBTE.

based data mining. We intend to study our model on other kinds of location-based data. For example, we can apply our model on location-based tweets from Twitter. We also plan to extend our model on other text topic modeling tasks. For example, we can conduct geographical sentiment analysis to seek users' personal interests. This is especially useful in determining the marketing strategy.

ACKNOWLEDGEMENT

This work is fully supported by a research funding from Google Focused Grant Project "Mobile 2014", and two grants from the Research Grants Council of the Hong Kong SAR, China (Project No. CUHK 413210 and Project No. CUHK 415410).

REFERENCES

1. C. E. Antoniak. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *Annals of Statistics*, 2(6):1152–1174, 1974.

2. L. Backstrom, E. Sun, and C. Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 61–70, New York, NY, USA, 2010. ACM.

3. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

4. J. Chon and H. Cha. Lifemap: A smartphone-based context provider for location-based services. *IEEE Pervasive Computing*, 10(2):58–67, 2011.

5. D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 761–770, New York, NY, USA, 2009. ACM.

6. S. Dhar and U. Varshney. Challenges and business models for mobile location-based services and advertising. *Commun. ACM*, 54(5):121–128, 2011.

7. N. Eagle, A. Pentland, and D. Lazer. Inferring social network structure using mobile phone data. *Proceedings of the National Academy of Sciences (PNAS)*, 106(36):15274–15278, 2009.

8. K. Farrahi and D. Gatica-Perez. Discovering routines from large-scale human locations using probabilistic topic models. *ACM TIST*, 2(1):3, 2011.

9. F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi. Trajectory pattern mining. In *KDD*, pages 330–339, 2007.

10. T. Hofmann. Probabilistic latent semantic analysis. In *UAI*, pages 289–296, 1999.

11. T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 50–57, New York, NY, USA, 1999. ACM.

12. Y. Liu, R. Yang, and E. Wilde. Open and decentralized access across location-based services. In *WWW (Companion Volume)*, pages 79–80, 2011.

DPM			LBTE			
Topic 1 Surfing	Topic 2 Hiking	Topic 3 Nature	Topic 1 Surfing	Topic 2 Hiking	Topic 3 Nature	Topic 4 Hiking
surfing 0.151154	hiking 0.205283	hiking 0.065217	surfing 0.085163	hiking 0.146915	nature 0.062500	hiking 0.161290
surf 0.106478	camping 0.194660	nature 0.063406	surf 0.083576	mountains 0.130261	yosemite 0.062500	arizona 0.134409
2009 0.096798	mountains 0.184324	walking 0.062953	surfer 0.083576	camping 0.082107	falls 0.062500	grandcanyon 0.134409
dog 0.094564	wyoming 0.181453	water 0.062500	water 0.083311	wyoming 0.081590	green 0.062500	trails 0.118280
hb 0.094564	cloudpeakwilderness 0.180879	park 0.062500	arm 0.083047	cloudpeakwilderness 0.081332	national 0.062500	southkaibabtrail 0.107527
spca 0.093820	washingtonisland 0.011484	rocks 0.062500	agua 0.083047	hike 0.042861	park 0.062500	trail 0.102151
n 0.093820	wisconsin 0.011484	yosemite 0.062047	oceano 0.083047	mikeonthetrail 0.040666	fresh 0.062500	unitedstates 0.102151
paws 0.093820	doorcounty 0.011484	green 0.062047	mar 0.083047	backpacking 0.035760	rocks 0.062500	usa 0.096774
ocsportsnut 0.017126	washington 0.005742	nationalpark 0.061594	miamibeach 0.083047	california 0.029305	boulders 0.062500	desert 0.032258
oceancitymaryland 0.017126	olympicnationalforest 0.005455	falls 0.061594	miami 0.083047	trail 0.028789	walking 0.062500	summer 0.005376

Table 5. Topics discovered in activities dataset from DMP and LBTE.

13. E. H.-C. Lu, V. S. Tseng, and P. S. Yu. Mining cluster-based temporal mobile sequential patterns in location-based service environments. *IEEE Trans. Knowl. Data Eng.*, 23(6):914–927, 2011.

14. Q. Mei, C. L. 0001, H. Su, and C. Zhai. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *WWW*, pages 533–542, 2006.

15. M. Meila. Comparing clusterings: an axiomatic view. In *ICML*, pages 577–584, 2005.

16. R. M. Neal. Markov chain sampling methods for dirichlet process mixture models. *Computational and Graphical Statistics*, 9:249–265, 2000.

17. B. K. Oksendal. *Stochastic Differential Equations: An Introduction with Applications*. Springer, 5 edition, 2002.

18. V. Rao and Y. W. Teh. Spatial normalized gamma processes. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1554–1562. 2009.

19. T. Rattenbury, N. Good, and M. Naaman. Towards automatic extraction of event and place semantics from flickr tags. In *SIGIR*, pages 103–110, 2007.

20. T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 851–860, New York, NY, USA, 2010. ACM.

21. S. Sizov. Geofolk: latent spatial semantics in web 2.0 social media. In *WSDM'10*, pages 281–290, 2010.

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

23. Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101:1566–1581, 2006.

24. C. Wang, J. Wang, X. Xie, and W.-Y. Ma. Mining geographic knowledge using location aware topic model. In *Proceedings of the 4th ACM workshop on Geographical information retrieval*, GIR '07, pages 65–70, New York, NY, USA, 2007. ACM.

25. Z. Yin, L. Cao, J. Han, C. Zhai, and T. Huang. Geographical topic discovery and comparison. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 247–256, New York, NY, USA, 2011. ACM.

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

27. C.-H. Yun and M.-S. Chen. Mining mobile sequential patterns in a mobile commerce environment. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 37(2):278–295, 2007.

28. Y. Zheng, L. Zhang, Z. Ma, X. Xie, and W.-Y. Ma. Recommending friends and locations based on individual location history. *TWEB*, 5(1):5, 2011.