Geographical Topic Discovery and Comparison

Zhijun Yin[1], Liangliang Cao[2], Jiawei Han[1], Chengxiang Zhai[1], Thomas Huang[2]
[1]Department of Computer Science
[2]Department of ECE and Beckman Institute
University of Illinois at Urbana-Champaign
zyin3@illinois.edu, cao4@ifp.uiuc.edu, hanj@cs.uiuc.edu, czhai@cs.uiuc.edu, huang@ifp.uiuc.edu

ABSTRACT

This paper studies the problem of discovering and comparing geographical topics from GPS-associated documents. GPS-associated documents become popular with the pervasiveness of location-acquisition technologies. For example, in Flickr, the geo-tagged photos are associated with tags and GPS locations. In Twitter, the locations of the tweets can be identified by the GPS locations from smart phones. Many interesting concepts, including cultures, scenes, and product sales, correspond to specialized geographical distributions. In this paper, we are interested in two questions: (1) how to discover different topics of interests that are coherent in geographical regions? (2) how to compare several topics across different geographical locations? To answer these questions, this paper proposes and compares three ways of modeling geographical topics: location-driven model, text-driven model, and a novel joint model called LGTA (Latent Geographical Topic Analysis) that combines location and text. To make a fair comparison, we collect several representative datasets from Flickr website including Landscape, Activity, Manhattan, National park, Festival, Car, and Food. The results show that the first two methods work in some datasets but fail in others. LGTA works well in all these datasets at not only finding regions of interests but also providing effective comparisons of the topics across different locations. The results confirm our hypothesis that the geographical distributions can help modeling topics, while topics provide important cues to group different geographical regions.

Categories and Subject Descriptors

H.2.8 [**Database applications**]: Data mining

General Terms

Algorithms

Keywords

Geographical topics, topic modeling, topic comparison

1. INTRODUCTION

With the popularity of low-cost GPS chips and smart phones, geographical records have become prevalent on the Web. A geographical record is usually denoted by a two dimensional vector, latitude and longitude, representing a unique location on the Earth. There are several popular ways to obtain geographical records on the Web:

1. Advanced cameras with GPS receivers could record GPS locations when the photos were taken. When users upload these photos on the Web, we can get the geographical records from the digital photo files.

2. Some applications including Google Earth and Flickr provide interfaces for users to specify a location on the world map. Such a location can be treated as a geographical record in a reasonable resolution.

3. People can record their locations by GPS functions in their smart phones. Popular social networking websites, including Facebook, Twitter, Foursquare and Dopplr, provide services for their users to publish such geographical information.

In the above three scenarios, GPS records are provided together with different documents including tags, user posts, etc. We name those documents with GPS records as *GPS-associated documents*. The amount of GPS-associated documents is increasing dramatically. For example, Flickr hosts more than 100 million photos associated with tags and GPS locations. The large amount of GPS-associated documents makes it possible to analyze the geographical characteristics of different subjects. For example, by analyzing the geographical distribution of food and festivals, we can compare the cultural differences around the world. We can also explore the hot topics regarding the candidates in presidential election in different places. Moreover, we can compare the popularity of specific products in different regions and help make the marketing strategy. The geographical characteristics of these topics call for effective approaches to study the GPS-associated documents on the Web.

In recent years, some studies have been conducted on GPS-associated documents including organizing geo-tagged photos [4] and searching large geographical datasets [7]. However, none of them addressed the following two needs in analyzing GPS-associated documents.

- *Discovering different topics of interests those are coherent in geographical regions.* Administrative divisions such as countries and states can be used as regions to discover topics. However, we are more interested in different region segmentations corresponding to different topics. For example, a city can be grouped into different sub-regions in terms of architecture or

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0632-4/11/03.

entertainment characteristics; a country might be separated into regions according to landscapes like desert, beach and mountain. Unfortunately, existing studies either overlook the differences across geographical regions or employ country/state as the fixed configuration.

- *Comparing several topics across different geographical locations.* It is often more interesting to compare several topics than to analyze a single topic. For example, people would like to know which products are more popular in different regions, and sociologists may want to know the cultural differences across different areas. With the help of GPS-associated documents, we can map topics of interests into their geographical distributions. None of the previous work addressed this problem and we aim to develop an effective method to compute such comparison.

In this paper, we propose three different models for geographical topic discovery and comparison. First, we introduce a location-driven model, where we cluster GPS-associated documents based on their locations and make each document cluster as one topic. The location-driven model works if there exist apparent location clusters. Second, we introduce a text-driven model, which discovers topics based on topic modeling with regularization by spatial information. The text-driven model can discover geographical topics if the regularizer is carefully selected. However, it cannot get the topic distribution in different locations for topic comparison, since locations are only used for regularization instead of being incorporated into the generative process. Third, considering the facts that a good geographical configuration benefits the estimation of topics, and that a good topic model helps identify the meaningful geographical segmentation, we build a unified model for both topic discovery and comparison. We propose a novel location-text joint model called LGTA (Latent Geographical Topic Analysis), which combines geographical clustering and topic modeling into one framework. Not only can we discover the geographical topics of high quality, but also can estimate the topic distribution in different geographical locations for topic comparison.

The rest of the paper is organized as follows. We formulate the problem of geographical topic discovery and comparison in Section 2. We introduce the location-driven model in Section 3 and the text-driven model in Section 4. In Section 5, we propose the Latent Geographical Topic Analysis model. We compare the performance of different methods in Section 6. We summarize the related work in Section 7 and conclude the paper in Section 8.

2. PROBLEM FORMULATION

In this section, we define the problem of geographical topic discovery and comparison. The notations used in this paper are listed in Table 1.

DEFINITION 1. A **GPS-associated document** is a text document associated with a GPS location. Formally, document d contains a set of words $\mathbf{w}_d$, where the words are from vocabulary set V. $l_d = (x_d, y_d)$ is the location of document d where x_d and y_d are longitude and latitude respectively. One example of a GPS-associated document can be a set of tags for a geo-tagged photo in Flickr, where the location is the GPS location where the photo was taken. Another example can be a tweet in Twitter, where the location is the GPS location from the smart phone.

Table 1: Notations used in the paper.

	Description
V	Vocabulary (word set), w is a word in V
D	Document collection
d	A document d that consists of words and GPS location
$\mathbf{w}_d$	The text of document d
l_d	The GPS location of document d
Z	The topic set, z is a topic in Z
θ	The word distribution set for Z, i.e., $\{\theta_z\}_{z \in Z}$

DEFINITION 2. A **geographical topic** is a spatially coherent meaningful theme. In other words, the words that are often close in space are clustered in a topic. We give two geographical topic examples as follows.

Example 1. Given a collection of geo-tagged photos related to **festival** with tags and locations in Flickr, the desired geographical topics are the festivals in different areas, such as *Cherry Blossom Festival* in Washington DC and *South by Southwest Festival* in Austin, etc.

Example 2. Given a collection of geo-tagged photos related to **landscape** with tags and locations in Flickr, the desired geographical topics are landscape categories that are spatially coherent, such as *coast*, *desert*, *mountain*, etc.

In this paper, we study the problem of geographical topic discovery and comparison. Given a collection of GPS-associated documents, we would like to discover the geographical topics. We would also like to compare the topics in different geographical locations. Here we give an example of geographical topic discovery and comparison.

Example 3. Given a collection of geo-tagged photos related to **food** with tags and locations in Flickr, we would like to discover the geographical topics, i.e., what people eat in different areas. After we discover the food preferences, we would like to compare the food preference distributions in different geographical locations.

To support topic comparison in different locations, we define the topic distribution in geographical location as follows.

DEFINITION 3. A **topic distribution in geographical location** is the conditional distribution of topics given a specific location. Formally, $p(z|l)$ is the probability of topic z given location $l = (x, y)$ where x is longitude and y is latitude, s.t., $\sum_{z \in Z} p(z|l) = 1$. From $p(z|l)$, we can know which topics are popular in location l.

The problem of **geographical topic discovery and comparison** is formulated as follows. Given a collection of GPS-associated documents D and the number of topics K, we would like to discover K geographical topics, i.e., $\theta = \{\theta_z\}_{z \in Z}$ where Z is the topic set and a geographical topic z is represented by a word distribution $\theta_z = \{p(w|z)\}_{w \in V}$ s.t. $\sum_{w \in V} p(w|z) = 1$. Along with the discovered geographical topics, we also would like to know the topic distribution in different geographical locations for topic comparison, i.e., $p(z|l)$ for all $z \in Z$ in location l as in Definition 3. In the next sections, we will present three different models for solving this problem.

3. LOCATION-DRIVEN MODEL

In the location-driven model, we simply cluster the documents based on their locations. Each document cluster corresponds to one topic. $p(z|d)$ is the probability of topic z given document d from the location clustering result. We then estimate the word distribution θ_z for topic z by $p(w|z) \propto \sum_{d \in D} p(w|d)p(d|z)$, where $p(d|z)$ is obtained from $p(z|d)$ by Bayes' theorem. In Festival dataset in Example 1, after we cluster the photos according to their locations, those photos close to each other are merged into the same cluster. And then we can generate the geographical topics (i.e., festival descriptions for each region) based on tags in each cluster.

To cluster objects in 2-D space, we can use partition-based clustering like KMeans, density-based clustering like Meanshift [3] and DBScan [5], and mixture model based clustering. After we get the word distribution θ_z for topic $z \in Z$ based on the clustering result, we would like to know the topic distribution in geographical location $p(z|l)$ for topic comparison. Therefore, we prefer a generative model for location clustering because we can get the estimation of $p(l|z)$. $p(z|l)$ can be obtained by Bayes' theorem from $p(l|z)$. A popular generative model is Gaussian Mixture Model (GMM). In GMM, we assume that each cluster is mathematically represented by a Gaussian distribution and the entire data set is modeled by a mixture of Gaussian distributions.

Although the location-driven model is straightforward, it is likely to fail if the document locations do not have good cluster patterns. A geographical topic may be from several different areas and these areas may not be close to each other. For example, in Landscape dataset in Example 2, there are no apparent location clusters; mountains exist in different areas and some are distant from each other. Therefore, the location-driven model fails in Landscape dataset as shown in the experiment in Section 6.2.1.

4. TEXT-DRIVEN MODEL

In the text-driven model, we discover the geographical topics based on topic modeling. To incorporate location information, we can use the idea of NetPLSA [8] to regularize topic modeling. PLSA [6] models the probability of each co-occurrence of words and documents as a mixture of conditionally independent multinomial distributions. NetPLSA regularizes PLSA with a harmonic regularizer based on a graph structure in the data. In our case, the nodes of the graph are documents and the edge weights are defined as the closeness in location between two documents. Therefore, documents that are close in location would be assumed to have similar topic distributions.

The objective function that NetPLSA aims to minimize is as follows.

$$L(D) = -(1-\lambda)\sum_{d\in D}\sum_{w\in V} c(w,d)\log\sum_{z\in Z} p(w|z)p(z|d) + \frac{\lambda}{2}\sum_{(u,v)\in E} w(u,v)\sum_{z\in Z}(p(z|d_u)-p(z|d_v))^2 \quad (1)$$

where $c(w,d)$ is the count of word w in document d and $w(u,v)$ is the closeness of document d_u and d_v. $p(w|z)$ is the word distribution of topic z and $p(z|d)$ is the topic distribution of document d. λ controls the regularization strength.

With the guidance of text information, the text-driven model may discover geographical topics that are missed by the location-driven model. However, there are still several problems in the text-driven model. First, we can only get the word distribution of geographical topics θ_z for $z \in Z$, but we cannot get the topic distribution of geographical locations in Definition 3, which is important for geographical topic comparison. In text-driven model we cannot know $p(z|l)$ because location is only used for regularization instead of being modeled in the topic generative process. Second, it is difficult to define the document closeness measure used in regularization. For example, in Food data set in Example 3, some food preferences exist only in some small regions, while some others exist throughout the continent. It is difficult to choose the closeness measure in this case.

Table 2: Notations used in LGTA framework.

	Description
R	The region set, r is a region in R
ϕ	The topic distribution set for R, i.e., $\{\phi_r\}_{r\in R}$
μ	The mean vector set for R, i.e., $\{\mu_r\}_{r\in R}$
Σ	The covariance matrix set for R, i.e., $\{\Sigma_r\}_{r\in R}$
α	The region importance weights

5. LOCATION-TEXT JOINT MODEL

In this section, we propose a novel location-text joint model called LGTA (Latent Geographical Topic Analysis), which combines geographical clustering and topic modeling into one framework.

5.1 General Idea

To discover geographical topics, we need a model to encode the spatial structure of words. The words that are close in space are likely to be clustered into the same geographical topic. In order to capture this property, we assume there are a set of regions. The topics are generated from regions instead of documents. 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. In Festival dataset in Example 1, the regions can be the areas in different cities, so the discovered geographical topics are different festivals. In Landscape data set in Example 2, the regions can be different areas such as the long strips along the coast and the areas in the mountains, so the discovered geographical topics are different landscapes. In Food data set in Example 3, the regions can be different areas that people live together, so the discovered geographical topics are different food preferences. We would like to design a model that can identify these regions as well as discover the geographical topics.

5.2 Latent Geographical Topic Analysis

In this section, we introduce our LGTA framework for geographical topic discovery and comparison. The notations used in the framework are listed in Table 2.

5.2.1 Discovering geographical topics

We would like to discover K geographical topics. The word distribution set of all the topics is denoted as θ, i.e., $\{\theta_z\}_{z\in Z}$. Let us assume there are N regions and denote the region set as R. We assume that the geographical distribution of each region is Gaussian, parameterized as $(\mu, \Sigma) = \{(\mu_r, \Sigma_r)\}_{r\in R}$ where μ_r and Σ_r are the mean vector and covariance matrix of region r. α is a weight distribution over all the regions. $p(r|\alpha)$ indicates the weight of region r and

$\sum_{r\in R} p(r|\alpha) = 1$. Since topics are generated from regions, we use $\phi = \{\phi_r\}_{r\in R}$ to indicate topic distributions for all the regions. $\phi_r = \{p(z|r)\}_{z\in Z}$ where $p(z|r)$ is the probability of topic z given region r. $\sum_{z\in Z} p(z|r) = 1$ for each r.

In our model, topics are generated from regions instead of documents and the geographical distribution of each region follows a Gaussian distribution. 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. The generative procedure of the model is described as follows.

To generate a geographical document d in collection D:

1. Sample a region r from the discrete distribution of region importance α, $r \sim Discrete(\alpha)$.
2. Sample location l_d from Gaussian distribution of μ_r and Σ_r.

$$p(l_d|\mu_r, \Sigma_r) = \frac{1}{2\pi\sqrt{|\Sigma_r|}} \exp(\frac{-(l_d-\mu_r)^T\Sigma_r^{-1}(l_d-\mu_r)}{2}) \tag{2}$$

3. To generate each word in document d:
   (a) Sample a topic z from multinomial ϕ_r.
   (b) Sample a word w from multinomial θ_z.

Instead of aligning each topic with a single region, each topic in our model can be related to several regions. Therefore, our model can handle topics with complex shapes. Our model identifies the regions considering both location and text information. Meanwhile, it discovers the geographical topics according to the identified geographical regions. Let us denote all parameters by $\Psi = \{\theta, \alpha, \phi, \mu, \Sigma\}$. Given the data collection $\{(\mathbf{w}_d, l_d)\}_{d\in D}$ where $\mathbf{w}_d$ is the text of document d and l_d is the location of document d, the log-likelihood of the collection given Ψ is as follows.

$$\begin{aligned} L(\Psi; D) &= \log p(D|\Psi) \\ &= \log \prod_{d\in D} p(\mathbf{w}_d, l_d|\Psi) \end{aligned} \tag{3}$$

In Section 5.3, we show how to estimate all the parameters using an EM algorithm.

5.2.2 Comparing geographical topics

To compare the topics in different geographical locations, we need to get $p(z|l)$ in Definition 3 for all topics $z \in Z$ given location $l = (x, y)$ where x is longitude and y is latitude. Given the estimated Ψ, we first estimate the density of location l given topic z.

$$\begin{aligned} p(l|z, \Psi) &= \sum_{r\in R} p(l|r, \Psi)p(r|z, \Psi) \\ &= \sum_{r\in R} p(l|\mu_r, \Sigma_r)\frac{p(z|r)p(r|\alpha)}{p(z|\Psi)} \end{aligned} \tag{4}$$

where $p(z|\Psi) = \sum_{r\in R} p(z|r)p(r|\alpha)$ and $p(l|\mu_r, \Sigma_r)$ is based on Equation 2.

After we get $p(l|z, \Psi)$, we can get $p(z|l, \Psi)$ according to Bayes' theorem.

$$\begin{aligned} p(z|l, \Psi) &\propto p(l|z, \Psi)p(z|\Psi) \\ &\propto \sum_{r\in R} p(l|\mu_r, \Sigma_r)p(z|r)p(r|\alpha) \end{aligned} \tag{5}$$

5.3 Parameter Estimation

In order to estimate parameters $\Psi = \{\theta, \alpha, \phi, \mu, \Sigma\}$ in Equation 3, we use maximum likelihood estimation. Specifically, we use Expectation Maximization(EM) algorithm to solve the problem, which iteratively computes a local maximum of likelihood. Let us denote r_d as the region of document d. We introduce the hidden variable $p(r|d, \Psi)$, which is the probability of $r_d = r$ given document d and Ψ. In the E-step, it computes the expectation of the complete likelihood $Q(\Psi|\Psi^{(t)})$, where $\Psi^{(t)}$ is the value of Ψ estimated in iteration t. In the M-step, it finds the estimation $\Psi^{(t+1)}$ that maximizes the expectation of the complete likelihood. The derivation detail is listed in Appendix A.

In the **E-step**, $p(r|d, \Psi^{(t)})$ is updated according to Bayes formulas as in Equation 6.

$$p(r|d, \Psi^{(t)}) = \frac{p^{(t)}(r|\alpha)p(\mathbf{w}_d, l_d|r, \Psi^{(t)})}{\sum_{r'\in R} p^{(t)}(r'|\alpha)p(\mathbf{w}_d, l_d|r', \Psi^{(t)})} \tag{6}$$

where $p(\mathbf{w}_d, l_d|r, \Psi^{(t)})$ is calculated as follows.

$$p(\mathbf{w}_d, l_d|r, \Psi^{(t)}) = p(\mathbf{w}_d|r, \Psi^{(t)})p(l_d|r, \Psi^{(t)}) \tag{7}$$

where $p(l_d|r, \Psi^{(t)}) = p(l_d|\mu_r^{(t)}, \Sigma_r^{(t)})$ is defined as Gaussian distribution in Equation 2 and $p(\mathbf{w}_d|r, \Psi^{(t)})$ is multinomial distribution for the words in document d in terms of probability $p(w|r, \Psi^{(t)})$.

$$p(\mathbf{w}_d|r, \Psi^{(t)}) \propto \prod_{w\in\mathbf{w}_d} p(w|r, \Psi^{(t)})^{c(w,d)} \tag{8}$$

where $c(d, w)$ is the count of word w in document d.

We assume that the words in each region are generated from a mixture of a background model and the region-based topic models. The purpose of using a background model is to make the topics concentrated more on more discriminative words, which leads to more informative models [16].

$$p(w|r, \Psi^{(t)}) = \lambda_B p(w|B) + (1-\lambda_B)\sum_{z\in Z} p^{(t)}(w|z)p^{(t)}(z|r) \tag{9}$$

$p^{(t)}(w|z)$ is from $\theta^{(t)}$, and $p^{(t)}(z|r)$ is from $\phi^{(t)}$. $p(w|B)$ is the background model, which we set as follows.

$$p(w|B) = \frac{\sum_{d\in D} c(w, d)}{\sum_{w\in V}\sum_{d\in D} c(w, d)} \tag{10}$$

In the **M-step**, we find the estimation $\Psi^{(t+1)}$ that maximizes the expectation of the complete likelihood $Q(\Psi|\Psi^{(t)})$ using the following updating formulas.

$$p^{(t+1)}(r|\alpha) = \frac{\sum_{d\in D} p(r|d, \Psi^{(t)})}{|D|} \tag{11}$$

$$\mu_r^{(t+1)} = \frac{\sum_{d\in D} p(r|d, \Psi^{(t)})l_d}{\sum_{d\in D} p(r|d, \Psi^{(t)})} \tag{12}$$

$$\Sigma_r^{(t+1)} = \frac{\sum_{d\in D} p(r|d, \Psi^{(t)})(l_d-\mu_r^{(t)})(l_d-\mu_r^{(t)})^T}{\sum_{d\in D} p(r|d, \Psi^{(t)})} \tag{13}$$

In order to get updated $\theta^{(t+1)}$ and $\phi^{(t+1)}$ in the M-step, we use another EM algorithm to estimate them. We define the hidden variable $\varphi(w, r, z)$, which corresponds to the events that word w in region r is from topic z. The relevant EM updating process is as follows.

$$\varphi(w,r,z) \leftarrow \frac{(1-\lambda_B)p(w|z)p(z|r)}{\lambda_B p(w|B) + (1-\lambda_B)\sum_{r\in R} p(w|z)p(z|r)} \quad (14)$$

$$p(z|r) \leftarrow \frac{\sum_{w\in V} c(w,d)p(r|d,\Psi^{(t)})\varphi(w,r,z)}{\sum_{z'\in Z}\sum_{w\in V} c(w,d)p(r|d,\Psi^{(t)})\varphi(w,r,z')} \quad (15)$$

$$p(w|z) \leftarrow \frac{\sum_{d\in D} c(w,d)p(r|d,\Psi^{(t)})\varphi(w,r,z)}{\sum_{w'\in V}\sum_{d\in D} c(w',d)p(r|d,\Psi^{(t)})\varphi(w',r,z)} \quad (16)$$

θ and ϕ obtained from the above EM steps are considered as $\theta^{(t+1)}$ and $\phi^{(t+1)}$.

5.4 Discussion

5.4.1 Complexity analysis

We analyze the complexity of parameter estimation process in Section 5.3. In the E-step, it needs $O(KN|V|)$ to calculate $p(w|r,\Psi^{(t)})$ in Equation 9 for all (w,r) pairs, where K is the number of topics, N is the number of regions and $|V|$ is the vocabulary size. To calculate $p(\mathbf{w}_d|r,\Psi^{(t)})$ in Equation 8 for all (d,r) pairs, it needs $O(N|W|)$ where $|W|$ is the total counts of the words in all the documents. It also needs $O(|D|)$ to calculate $p(l_d|r,\Psi^{(t)})$ for all the documents. Therefore, the complexity of getting $p(r|d,\Psi^{(t)})$ for all (r,d) pairs is $O(KN|V|+N|W|)$. In the M-step, it needs $O(N|D|)$ to get the updated $p^{(t+1)}(r|\alpha)$, $\mu_r^{(t+1)}$ and $\Sigma_r^{(t+1)}$ as in Equations 11, 12 and 13 for all the regions. To get updated $\theta^{(t+1)}$ and $\phi^{(t+1)}$, it needs $O(T_2KN|V|)$ where T_2 is the number of iterations for Equations 14, 15 and 16. Therefore, the complexity of M-step is $O(N|D| + T_2KN|V|)$. The complexity of the whole framework is $O(T_1(KN|V| + N|W| + N|D| + T_2KN|V|))$, where T_1 is the number of iterations in the EM algorithm.

5.4.2 Parameter setting

In our model, we have three parameters, i.e., the mixing weight of the background model λ_B, the number of topics K and the number of regions N. A large λ_B can exclude the common words from the topics. In this paper λ_B is fixed as 0.9 following the empirical studies [16, 9]. K is the desired number of geographical topics. Users can specify the value of K according to their needs. N is the number of the regions used in our model for generating the topics, which provides the flexility for users to adjust the granularity of regions. The larger N is, the more fine-grained the regions are. For example, in Landscape dataset in Example 2, a large N is preferred, since we would like to use fine-grained regions to handle complex shapes of different landscape categories. In Festival dataset in Example 1, N is preferred to be close to K, since we would like to discover the topics in different areas. In our experiment, small changes of N yield similar results. When the parameters are unknown, Schwarz's Bayesian information criterion (BIC) provides an efficient way to select the parameters. The BIC measure includes two parts: the log-likelihood and the model complexity. The first part characterizes the fitness over the observations, while the second is determined by the number of parameters. In practice we can train models with different parameters, and compare their BIC values. The model with the lowest value will be selected as the final model.

5.4.3 Topic guidance in comparison

We can add some guidance in the framework to make the discovered geographical topics aligned with our needs for topic comparison. For example, in Food data set, we would like to compare the geographical distribution of *Chinese* food and *Italian* food, we can add some prior knowledge in two topics and guide one topic to be related to *Chinese* food and the other to be related to *Italian* food. Specifically, we define a conjugate prior (i.e., Dirichlet prior) on each multinomial topic distribution. Let us denote the Dirichlet prior σ_z for topic z. $\sigma_z(w)$ can be interpreted as the corresponding pseudo counts for word w when we estimate the topic distribution $p(w|z)$. With this conjugate prior, we can use the Maximum a Posteriori (MAP) estimator for parameter estimation, which can be computed using the same EM algorithm except that we would replace Equation 16 with the following formula:

$$p(w|z) \leftarrow \frac{\sum_{d\in D} c(w,d)p(r|d,\Psi^{(t)})\varphi(w,r,z) + \sigma_z(w)}{\sum_{w'\in V}\sum_{d\in D}(c(w',d)p(r|d,\Psi^{(t)})\varphi(w',r,z) + \sigma_z(w'))} \quad (17)$$

5.4.4 Comparison with GeoFolk

In [13], Sizov proposed a novel model named GeoFolk to combine the semantics of text feature and spatial knowledge. Sizov shows that GeoFolk works better than text-only analysis in tag recommendation, content classification and clustering. However, GeoFolk is not suitable for region clustering due to two facts: First, GeoFolk models each region as an isolated topic and thus fails to find the common topics in different geographical sites. Second, GeoFolk assumes the geographical distribution of each topic is Gaussian, which makes its results similar to the results of the location-driven model using GMM. As a result, it would fail to discover the meaningful topics with non-Gaussian geographical distributions. For example, in the Landscape dataset in Example 2, the *coast* topic is along the coastline, GeoFolk fails to discover it. For the *mountain* topic, GeoFolk cannot discover it because the *mountain* topic is located in different areas. In contrast, our LGTA model separates the concepts of topics and regions, and the coordinates are generated from regions instead of topics. Therefore, we can discover the meaningful geographical topics properly.

6. EXPERIMENT

6.1 Data Set

We evaluate the proposed models on Flickr dataset. We crawl the images with GPS locations through Flickr API [1]. Flickr API supports search criteria including tag, time, GPS range, etc.. We select several representative topics including Landscape, Activity, Manhattan, National Park, Festival, Car and Food. The statistics of the datasets are listed in Table 3. For Landscape dataset, we crawl the images containing tag *landscape* and keep the images containing tags *mountains, mountain, beach, ocean, coast, desert* around US. For Activity data set, we crawl the images containing tags *hiking* and *surfing* around US. For Manhattan dataset, we crawl the images containing tag *manhattan* in New York City. For National Park dataset, we crawl the images containing

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

Table 3: The statistics of the datasets.

Data set	Time span	# image	# words
Landscape	09/01/09 - 09/01/10	5791	1143
Activity	09/01/09 - 09/01/10	1931	408
Manhattan	09/01/09 - 09/01/10	28922	868
Festival	09/01/09 - 09/01/10	1751	421
National Park	09/01/09 - 09/01/10	2384	351
Car	01/01/06 - 09/01/10	34707	12
Food	01/01/06 - 09/01/10	151747	278

tag *nationalpark* and keep the images with tags *rockymountain, yellowstone, olympic, grandcanyon, everglades, smokymountain, yosemite, acadia*. For Festival dataset, we crawl the images containing tag *festival* in New York, Los Angeles, Chicago, Washington DC, San Francisco and Austin area. For Car data set, we crawl the images containing tags *chevrolet, pontiac, cadillac, gmc, buick, audi, bmw, mercedesbenz, fiat, peugeot, citroen, renault*. We remove the images with tags *autoshow, show, race, racing* and only keep car brand names in the dataset. For Food dataset, we crawl the images containing tags *cuisine, food, gourmet, restaurant, restaurants, breakfast, lunch, dinner, appetizer, entree, dessert* and keep 278 related food tags including dish names and food style names.

We compare the following methods in the experiment.

- LDM: Location-driven model in Section 3.
- TDM: Text-driven model in Section 4. We set regularization factor λ as 0.5 and add one edge between two documents if their distance is within threshold ε. ε varies according to different settings in the datasets as shown in Section 6.2.
- GeoFolk: The topic modeling method proposed in [13], which uses both text and spatial information (see Section 5.4.4).
- LGTA: Latent Geographical Topic Analysis framework in Section 5.

6.2 Geographical Topic Discovery

In this section, we compare the discovered geographical topics by different methods in several representative datasets.

6.2.1 Topic discovery for Landscape dataset

In Landscape dataset, we intend to discover 3 topics, i.e., different landscapes. We set ε in TDM as 0.1(∼10km), since we assume that two locations within 10km should have similar landscapes. In LGTA, we set the number of regions N as 30, since we would like to use 10 regions in average to cover each landscape topic. We list the topics discovered by different methods in Table 4, and we also plot the document locations for different topics on the map in Figure 1. Since there are no apparent location clusters for the topics, LDM and GeoFolk fail to discover meaningful geographical topics due to their inappropriate assumption that each topic has a location distribution like Gaussian. TDM performs better than LDM and GeoFolk. Topic 1 of TDM is related to *coast*, but Topic 2 and Topic 3 are not distinguishable. In LGTA, we assume that the topics are generated from a set of regions, so we can clearly identify three clusters *coast*, *desert* and *mountain* in Table 4. From the LGTA topics in Figure 1, we can see that Topic 1(*coast*) is along the coastline, Topic 2(*desert*) is aligned with the desert areas in US and Topic 3(*mountain*) maps to the mountain areas in US.

6.2.2 Topic discovery for Activity dataset

In Activity dataset, we intend to discover 2 topics, i.e., *hiking* and *surfing*. We set ε in TDM as 0.1(∼10km), since we assume that two locations within 10km should have similar activities. In LGTA, we set the number of regions N as 20, since we would like to use 10 regions in average to cover each activity topic. Similar to Landscape dataset, LDM and GeoFolk fail to discover meaningful geographical topics because there are no apparent location clusters for the topics. The result of LDM is similar to GeoFolk, while the result of TDM is similar to LGTA. Both TDM and LGTA can identify two topics, i.e., *hiking* and *surfing*. We list the topics discovered by GeoFolk and LGTA in Table 5.

Table 5: Topics discovered for Activity dataset.

GeoFolk		LGTA	
Topic 1	Topic 2	Topic 1(surfing)	Topic 2(hiking)
hiking 0.077	hiking 0.095	surfing 0.070	hiking 0.109
mountains 0.037	mountains 0.050	beach 0.065	mountains 0.059
mountain 0.027	mountain 0.041	california 0.059	mountain 0.042
california 0.027	surfing 0.032	ocean 0.053	nature 0.027
surfing 0.024	beach 0.030	surf 0.031	trail 0.019
beach 0.023	[nh] 0.029	hiking 0.031	hike 0.017
nature 0.020	white[mtn]s 0.022	waves 0.028	desert 0.017
ocean 0.019	trail 0.021	water 0.025	washington 0.014
trail 0.015	ocean 0.021	surfer 0.022	lake 0.013
hike 0.015	nature 0.019	pacific 0.018	camping 0.013

*[mtn] is mountain. [nh] is newhampshire.

6.2.3 Topic discovery for Manhattan dataset

In Manhattan dataset, we intend to discover 5 topics, i.e., different regions in Manhattan. We set ε in TDM as 0.001(∼0.1km), since the photos in Manhattan are very dense. In LGTA, we make the number of regions close to the number of topics, since we would like to discover large regions in Manhattan. We set the number of regions N as 10. Overall, LDM, GeoFolk and LGTA can identify different regions in Manhattan because meaningful topics can be obtained by clustering based on location, such as topic *lowermanhattan* and topic *midtown*. Although we have the regularization based on spatial information in TDM, it can only guarantee the smoothness of topics in the neighborhood. TDM is likely to mix the words from distant areas in the same topic. For example, TDM mix *timessquare 0.060*, *upperwestside 0.051*, *chinatown 0.033*, *greenwichvillage 0.031* and *unionsquare 0.017* into one topic, and these words are distant from each other.

6.2.4 Topic discovery for Festival dataset

In Festival dataset, we intend to discover 10 topics, i.e., festivals in different cities. We set ε in TDM as 0.01(∼1km), since 1km is a reasonable range in cities. In LGTA, we set the number of regions N as 20. Similar to Manhattan dataset, LDM, GeoFolk and LGTA can discover meaningful geographical topics, because the cities are distant from each other in space. TDM is possible to mix the festivals from different areas into the same topic. We list the topics related to *southbysouthwest* festival discovered by TDM, GeoFolk and LGTA in Table 6. The result of LDM is similar to GeoFolk. From Table 6, we can find that GeoFolk and LGTA discover pure topics related to *southbysouthwest* festival in Austin, but TDM mix *southbysouthwest* in Austin and *atlanticantic streetfair* in New York together.

6.2.5 Topic discovery for National Park dataset

In National Park dataset, we intend to discover 8 topics, i.e., different national parks. We set ε in TDM as 0.01(∼1km), since 1km is a reasonable range in national park

Table 4: Topics discovered for Landscape dataset.

LDM			TDM			GeoFolk			LGTA		
Topic 1	Topic 2	Topic 3	Topic 1	Topic 2	Topic 3	Topic 1	Topic 2	Topic 3	Topic 1	Topic 2	Topic 3
california	mountains	beach	ocean	mountains	mountains	california	desert	beach	beach	desert	mountains
ocean	desert	ocean	beach	desert	water	ocean	mountains	ocean	ocean	california	mountain
mountains	mountain	water	california	mountain	mountain	water	mountain	water	water	mountains	lake
water	utah	mountains	water	california	trees	beach	california	mountains	california	mountain	trees
beach	arizona	sea	sea	utah	coast	mountains	water	sea	sea	arizona	water
desert	lake	sunset	sunset	nationalpark	lake	coast	utah	sunset	coast	utah	snow
mountain	snow	mountain	seascape	snow	reflection	mountain	arizona	mountain	sunset	rock	scenery
sunset	southwest	blue	sand	rock	oregon	sea	sunset	blue	seascape	southwest	hiking
coast	rock	seascape	arizona	park	scenery	sunset	rock	seascape	pacific	park	washington
sea	water	lake	blue	lake	washington	pacific	snow	lake	sand	sunset	reflection

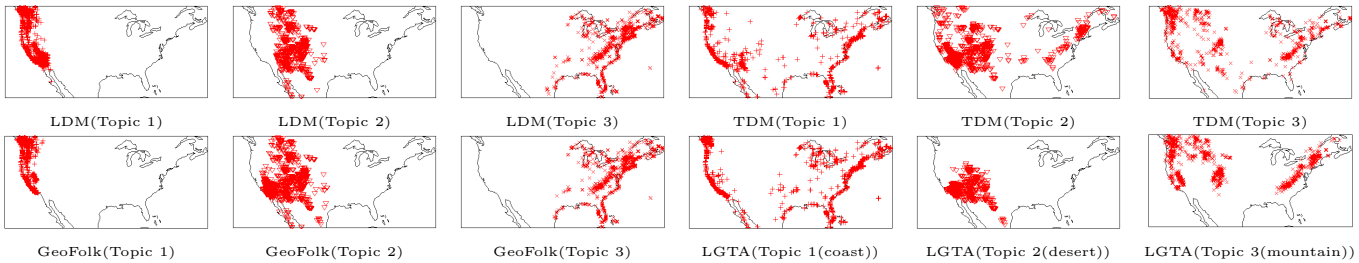


Figure 1: The document locations of different topics for Landscape dataset.

Table 6: Topic *southbysouthwest* for Festival dataset.

TDM	GeoFolk	LGTA
sxsw 0.124	sxsw 0.173	sxsw 0.163
brooklyn 0.082	austin 0.136	austin 0.149
southbysouthwest 0.061	southbysouthwest 0.127	texas 0.142
south 0.055	texas 0.125	southbysouthwest 0.085
streetfestival 0.050	south 0.121	south 0.070
southwest 0.049	southwest 0.103	funfunfunfest 0.061
funfunfunfest 0.044	downtown 0.093	southwest 0.060
atlanticavenue 0.044	musicfestival 0.074	musicfestival 0.057
atlanticantic 0.041	live 0.034	downtown 0.040
streetfair 0.040	stage 0.010	music 0.034

areas. In LGTA, we set the number of regions N as 20. We show that even if there are apparent location clusters, LDM and GeoFolk may obtain misleading results. As shown in Table 7, GeoFolk merges *acadia*, *everglades* and *greatsmoky-mountain* together into topic *acadia*, because these three national parks have fewer photos than other parks and are all located on the east coast of US. GeoFolk, similar to LDM, uses one Gaussian distribution to cover all these three parks, so the words from these parks are mixed into a single topic. In TDM, topic *acadia* is mixed with *rockymountain*. In LGTA, we use the fine-grained regions to generate the topics, so all the words in LGTA are related to *acadia*, where *mountdesertisland* is home to *acadia* and *barharbor* is a town on *mountdesertisland*.

Table 7: Topic *acadia* for National park dataset.

TDM	GeoFolk	LGTA
acadia[npk] 0.088	acadia[npk] 0.108	acadia[npk] 0.208
maine 0.087	maine 0.107	maine 0.205
acadia 0.087	acadia 0.107	acadia 0.205
colorado 0.081	everglades 0.079	barharbor 0.084
rocky[mtn][npk] 0.071	florida 0.058	newengland 0.084
northrim 0.050	tennessee 0.050	mountdesert[isl] 0.070
rockymountain 0.036	barharbor 0.043	beach 0.025
newengland 0.036	newengland 0.043	outdoor 0.016
barharbor 0.036	greatsmoky[mtn][npk] 0.043	flowers 0.015
rockymountains 0.034	mountdesert[isl] 0.036	wood 0.012

*[mtn] is mountain. [npk] is nationalpark. [isl] is island.

6.2.6 Topic discovery for Car dataset.

In Car dataset, we intend to discover 3 topics. We set ε in TDM as 0.1(∼10km), since 10km is a reasonable range in the world scale. In LGTA, we would like to use the fine-grained regions to discover the possible topics, so we set the number of regions N as 50. In Car dataset, there are no apparent location clusters or good text indications. As shown in Table 8, LDM, TDM and GeoFolk all fail to discover meaningful topics. However, LGTA can get the interesting geographical topics. In LGTA, Topic 1 is about American cars including *chevrolet*, *pontiac*, *cadillac*, *gmc* and *buick*. Topic 2 is related to German cars including *audi*, *mercedesbenz* and *bmw*. Topic 3 is about those European cars excluding German brands, including *fiat*, *peugeot*, *citroen* and *renault*. These interesting patterns can be discovered because these car brands in the same topic have similar geographical distributions.

6.2.7 Summary

With the experiments on these representative datasets, we can summarize the results as follows. If there are apparent location cluster patterns such as Manhattan and Festival datasets, LDM and GeoFolk are able to work, so is LGTA. If there are no apparent location clusters but good text indications in the datasets such as Landscape and Activity datasets, LDM and GeoFolk fail, TDM may work and LGTA works well. Even if there are location cluster patterns, LDM and GeoFolk may fail, while LGTA is still robust, such as in National Park dataset. In the difficult datasets such as Car dataset, only LGTA can discover meaningful geographical topics. Overall, LGTA is the best and most robust method for geographical topic discovery.

6.3 Quantitative Measures

In this section, we use some quantitative measures to evaluate the performances of different methods.

We use perplexity to evaluate the performance of topic modeling [1]. We keep 80% of the data collection as the train set and use the remaining collection as the held-out test set. We train the models on the train set and compute the perplexity of the test set to evaluate the models. A lower perplexity score indicates better generalization performance of the model. Specifically, we use text perplexity to measure the topic qualities and use location/text perplexity to measure the performance of geographical topics.

$$perplexity_{text}(D_{test}) = exp\{-\frac{\sum_{d\in D_{test}} \log p(\mathbf{w}_d)}{\sum_{d\in D_{test}} N_d}\}$$

$$perplexity_{location/text}(D_{test}) = exp\{-\frac{\sum_{d\in D_{test}} \log p(\mathbf{w}_d, l_d)}{\sum_{d\in D_{test}} N_d}\}$$

Table 8: Topics discovered for Car dataset

LDM			TDM			GeoFolk			LGTA		
Topic 1	Topic 2	Topic 3	Topic 1	Topic 2	Topic 3	Topic 1	Topic 2	Topic 3	Topic 1	Topic 2	Topic 3
chevrolet	chevrolet	fiat	bmw	renault	cadillac	fiat	peugeot	chevrolet	fiat	bmw	chevrolet
gmc	pontiac	renault	chevrolet	peugeot	audi	renault	chevrolet	pontiac	renault	audi	pontiac
cadillac	cadillac	citroen	fiat	mercedesbenz	pontiac	citroen	bmw	cadillac	citroen	mercedesbenz	cadillac
buick	buick	peugeot	citroen	buick	gmc	peugeot	fiat	gmc	peugeot	-	gmc
pontiac	gmc	audi	buick	-	buick	mercedesbenz	renault	buick	-	-	buick

*If the probability of a word in a topic is less than 1e-4, output as '-'.

where D_{test} is the test collection and N_d is document length of document d.

We list the results of text perplexity for different methods in Table 9 and the results of location/text perplexity for LDM, GeoFolk and LGTA in Table 10. TDM is not available in Table 10 because we cannot estimate the location probabilities using TDM. From Table 9 and 10, we can see both text perplexity and location/text perplexity of LGTA are the lowest in all the datasets. Especially, in Landscape, Activity and Car datasets, neither LDM nor GeoFolk can discover meaningful geographical topics, so the perplexities of LDM and GeoFolk in these data sets are much larger than those of LGTA.

Table 9: Text perplexity in datasets.

Data set	LDM	TDM	GeoFolk	LGTA
Landscape	394.680	444.676	384.411	**366.546**
Activity	184.970	176.234	184.979	**157.775**
Manhattan	193.823	201.042	193.001	**192.010**
National Park	118.159	120.100	117.238	**117.077**
Festival	177.978	214.975	173.621	**170.033**
Car	9.936	9.926	9.937	**9.924**

Table 10: Location/text perplexity in datasets.

Data set	LDM	GeoFolk	LGTA
Landscape	688.628	672.967	**569.047**
Activity	358.559	358.577	**257.086**
Manhattan	109.103	107.620	**105.684**
National Park	136.435	112.973	**103.853**
Festival	99.308	94.604	**91.230**
Car	40242.767	40348.974	**8718.927**

In Table 11, we show the average distance of word distributions of all pairs of topics measured by KL-divergence. The larger the average KL-divergence is, the more distinct the topics are. In Landscape and Activity datasets, LDM and GeoFolk fail to discover meaningful topics, so the average KL-divergence of TDM and LGTA is much larger than those of LDM and GeoFolk. In Manhattan, National Park and Festival datasets, the average KL-divergence of different methods are similar. In Car datasets, the average KL-divergence of TDM and LGTA are much larger than LDM and GeoFolk. Although the words from different topics of TDM in Car dataset are distinct, the topics are not meaningful as shown in Section 6.2.6.

Table 11: Average KL-divergence between topics in datasets.

Data set	LDM	TDM	GeoFolk	LGTA
Landscape	0.159	**0.311**	0.141	0.281
Activity	0.164	0.402	0.164	**0.491**
Manhattan	0.908	**1.091**	0.965	1.020
National Park	2.576	2.325	2.474	**2.598**
Festival	2.206	2.109	2.080	**2.258**
Car	2.518	**3.745**	2.365	3.731

6.4 Geographical Topic Comparison

In this section, we show the results of topic comparison for Car and Food datasets.

6.4.1 Topic comparison for Car dataset

In Figure 2, we plot the topic distribution in different locations for Car dataset according to the discovered topics from LGTA in Section 6.2.6. Compared with European cars, American cars are mainly in North America. European excluding German cars dominate most of European areas. German cars, as luxury brands, are popular in Germany and other areas such as East Asia and Australia.

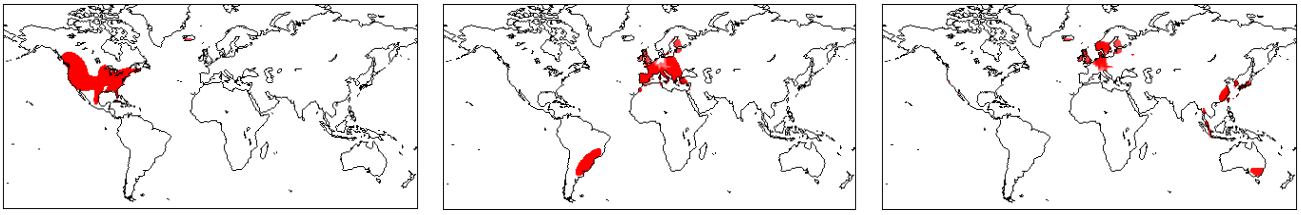
6.4.2 Topic comparison for Food dataset

In Food dataset, we set the number of topics K as 10. To derive the topics that we are interested in, we set the priors according to Equation 17. We use the words *chinese*, *japanese*, *italian*, *french*, *spanish* and *mexican* as priors for six topics and leave the remaining four topics to other possible food preferences. We set the number of regions N as 100, since we would like to use more find-grained regions to discover the food preferences. As shown in Table 12, each of the six topics consists of the typical food related to the preferences. We plot the comparison of the topics on the maps in Figure 3. From Figure 3, we can find that Chinese food is popular in China and Southeast Asia. In US and West Europe, Chinese food also has certain popularity. Japanese food is dominant in Japan, and it is welcome on the west coast of US. Italian food is very popular in Mediterranean area, and it is popular in US too. French food is popular in France and US. Spanish food is popular in Spain, US and part of South America. Mexican food is the main food in Mexico, and it highly influences the Southwestern area of US. From all these figures, we can find that each food preference has its main area. In the metropolitan areas in US, different kinds of food co-exist.

7. RELATED WORK

In this section we discuss some work related to our study, including geo-tagged social media mining, topic modeling and image processing using spatial coherence.

Geo-tagged social media mining With the development of GPS technology, several studies have been done in geo-tagged social media mining. Rattenbury et al. [11] use Scale-structure Identification method to extract place and event semantics for tags based on the GPS metadata of the images in Flickr. Crandall et al. [4] combine content analysis based on text tags and image data with structural analysis based on geospatial data to estimate the photo locations. In [7], Kennedy et al. use location, tags and visual features of the images to generate diverse and representative images for the landmarks. All these studies are related to the interplay between tags and locations in different applications, but they do not touch the problem of geographical topics discussed in this paper. Sizov [13] proposed a framework called GeoFolk to combine text and spatial information together to construct better algorithms for content management, retrieval, and sharing in social media. To make use of spatial information, GeoFolk assumes that each topic generates latitude and longitude from two topic-specific Gaussian



American Car European(excluding German) Car German Car

Figure 2: Topic comparison for Car dataset. For topic z, we plot $p(z|l)$ for all the locations. The larger $p(z|l)$ is, the darker the location is. We only plot the locations with $p(l|z) > 1e^{-4}$.

Table 12: Topic discovered for Food dataset.

Chinese Food	Japanese Food	Italian Food	French Food	Spanish Food	Mexican Food
chinese 0.552	japanese 0.519	italian 0.848	french 0.564	spanish 0.488	mexican 0.484
noodles 0.067	ramen 0.104	cappuccino 0.067	bistro 0.070	tapas 0.269	tacos 0.069
dimsum 0.064	soba 0.066	latte 0.048	patisserie 0.056	paella 0.076	taco 0.059
hotpot 0.039	noodle 0.065	gelato 0.030	bakery 0.049	pescado 0.059	salsa 0.036
rice 0.038	sashimi 0.039	pizza 0.002	resto 0.044	olives 0.032	cajun 0.031
noodle 0.035	yakitori 0.030	pizzeria 0.002	pastry 0.033	stickyrice 0.017	burrito 0.027
tofu 0.020	okonomiyaki 0.026	mozzarella 0.001	tarte 0.026	tortilla 0.013	crawfish 0.023
dumpling 0.018	udon 0.026	pasta 0.001	croissant 0.021	mediterranean 0.010	guacamole 0.022
duck 0.018	tempura 0.020	ravioli 0.000	baguette 0.019	mussels 0.008	margarita 0.020
prawn 0.017	curry 0.016	pesto 0.000	mediterranean 0.018	octopus 0.008	cocktails 0.020

distributions. However, geographical topics may not be like Gaussian distributions, such as topics "hiking" and "surfing". In our model, we distinguish the concepts of topics and regions and provide a more systematic way to discover geographical topics and we also provide geographical topic comparison which is not available in the existing models.

Topic modeling Topic modeling is a classic problem in text mining. The most representative models include PLSA [6] and LDA [1]. Wang et al. [15] use an LDA-style topic model to capture both the topic structure and the changes over time. In these studies, they do not consider the location information of the documents, so they do not focus on geographical topics. In [14], Wang et al. propose a Location Aware Topic Model to explicitly model the relationships between locations and words, where the locations are represented by predefined location terms in the documents. Mei et al.[9] proposed a probabilistic approach to model the subtopic themes and spatiotemporal theme patterns simultaneously in weblogs, where the locations need to be predefined. However, in geographical topic discovery, we do not know the locations or regions of interest beforehand. If we directly use the administrative region partitions, it would be difficult to discover topics whose corresponding regions are not aligned well with the pre-segmented regions. In [8], Mei et al. proposed a model called NetPLSA to combine PLSA with a graph-based regularizer, where adjacent nodes in document similarity graph should have similar topic distribution. We use NetPLSA in the text-driven model. However, NetPLSA cannot provide the geographical distribution of the topics. As shown in experiment, our LGTA model not only is more robust but also can provide interesting topic comparison results.

Spatial coherence inside images Our work is also partially motivated by the recent work in computer vision [2, 10, 12, 14] which try to simultaneously do object classification and segmentation in images. However, these studies are fundamentally different from this paper in three aspects. First, a spatial coherent segment is part of a image, while our geographical region contains multiple documents. This fundamental difference leads to different generative models. Second, segmentations in one image are usually clearly separated by contours and boundaries, which makes it possible to rely on superpixels [2, 12] to merge into an object of interests. However, there are no contours in geographical distribution. At last, the computer vision community focus on image classification instead of topic comparison, while the latter is important in Web mining.

8. CONCLUSION

The emerging trend of GPS-associated document opens up a wide variety of novel applications. In this paper, we introduce the problem of geographical topic discovery and comparison. We propose and compare three strategies of modeling geographical topics including location-driven model, text-driven model, and a novel joint model called LGTA (Latent Geographical Topic Analysis) that combines both location and text information. To test our approaches, we collect several representative datasets from Flickr website including Landscape, Activity, Manhattan, National park, Festival, Car, and Food. Evaluation results show that the new LGTA model works well for not only finding regions of interests but also providing effective comparisons of different topics across locations.

Our work opens up several interesting future directions. First, we can 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, other than topic discovery and comparison, we would like to extend our model to other text mining tasks. For example, we can do geographical sentiment analysis for different subjects.

Acknowledgement

Research was sponsored in part by the U.S. National Science Foundation under grants CCF-0905014, CNS-0931975, CNS-1027965, and IIS-0713581, by the Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053 (NS-CTA), and Air Force Office of Scientific Research MURI award FA9550-08-1-0265, in part by a Beckman Institute (Illinois) Seed Grant and in part by an NSF Grant IIS 1049332 EAGER. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

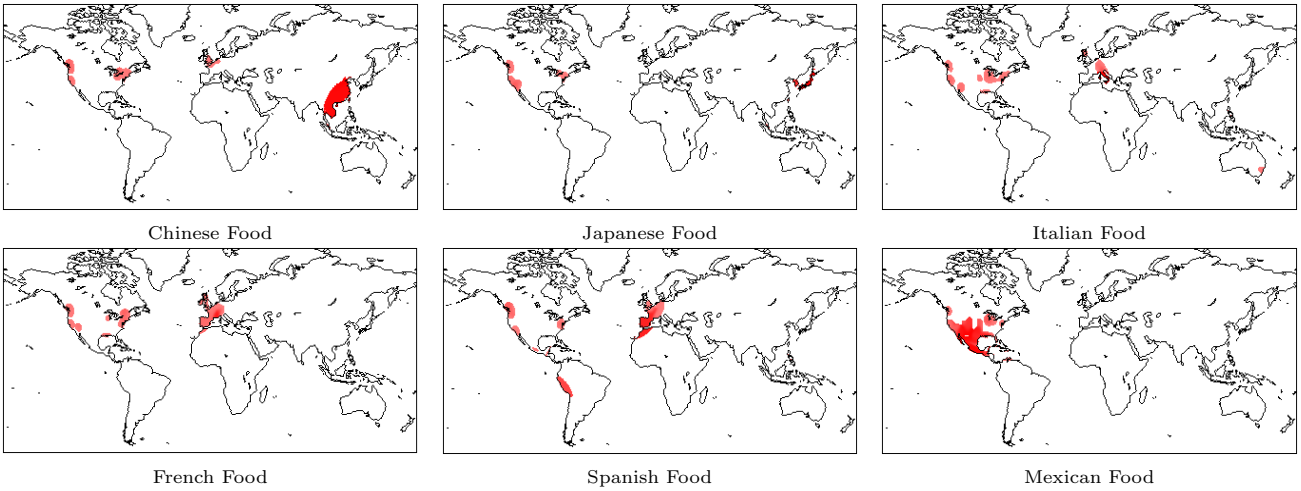


Figure 3: Topic comparison for Food dataset. For topic z, we plot $p(z|l)$ for all the locations. The larger $p(z|l)$ is, the darker the location is. We only plot the locations with $p(l|z) > 1e^{-4}$.

9. REFERENCES

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

[2] L. Cao and L. Fei-Fei. Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In *ICCV*, pages 1–8, 2007.

[3] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(5):603–619, 2002.

[4] D. J. Crandall, L. Backstrom, D. P. Huttenlocher, and J. M. Kleinberg. Mapping the world's photos. In *WWW*, pages 761–770, 2009.

[5] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, pages 226–231, 1996.

[6] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR*, pages 50–57, 1999.

[7] L. S. Kennedy and M. Naaman. Generating diverse and representative image search results for landmarks. In *WWW*, pages 297–306, 2008.

[8] Q. Mei, D. Cai, D. Zhang, and C. Zhai. Topic modeling with network regularization. In *WWW*, pages 101–110, 2008.

[9] Q. Mei, C. Liu, H. Su, and C. Zhai. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *WWW*, pages 533–542, 2006.

[10] J. C. Niebles, H. Wang, and F.-F. Li. Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision*, 79(3):299–318, 2008.

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

[12] B. C. Russell, W. T. Freeman, A. A. Efros, J. Sivic, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *CVPR (2)*, pages 1605–1614, 2006.

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

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

[15] X. Wang and A. McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *KDD*, pages 424–433, 2006.

[16] C. Zhai, A. Velivelli, and B. Yu. A cross-collection mixture model for comparative text mining. In *KDD*, pages 743–748, 2004.

APPENDIX

A. EM ALGORITHM DERIVATION

The expectation of the complete likelihood $Q(\Psi|\Psi^{(t)})$ is.

$$\begin{aligned} Q(\Psi|\Psi^{(t)}) &= E[\log \prod_{d\in D} p(r_d|\alpha)p(\mathbf{w}_d, l_d|r_d, \Psi)|D, \Psi^{(t)}] \\ &= \sum_{d\in D}\sum_{r\in R} p(r|d,\Psi^{(t)}) \log p(r|\alpha) + \\ &\quad \sum_{d\in D}\sum_{r\in R} p(r|d,\Psi^{(t)}) \log p(l_d|r,\Psi) + \\ &\quad \sum_{d\in D}\sum_{r\in R} p(r|d,\Psi^{(t)}) \log p(\mathbf{w}_d|r,\Psi) \end{aligned} \quad (18)$$

In the E-step, $p(r|d,\Psi^{(t)})$ is updated according to Bayes' rule as in Equation 6. In the M-step, it find the estimation $\Psi^{(t+1)}$ that maximize the complete likelihood $Q(\Psi|\Psi^{(t)})$. Since θ, α, ϕ, μ and Σ are in three different summands in Equation 18, we can optimize each summand separately to find $\Psi^{(t+1)}$ that maximize the complete likelihood $Q(\Psi|\Psi^{(t)})$ as in Equation 11, 12 and 13.

We use EM algorithm to find the optimal parameters θ and ϕ which maximize the last summand in Equation 18. The corresponding log-likelihood that we would like to maximize is as follows.

$$\begin{aligned} L(\theta,\phi;D) &= \sum_{d\in D}\sum_{r\in R}\sum_{w\in V} c(w,d)p(r|d,\Psi^{(t)}) \\ &\quad \log(\lambda_B p(w|B) + (1-\lambda_B)\sum_{z\in Z} p(w|z)p(z|r)) \end{aligned}$$

Equations 14, 15 and 16 correspond to the EM steps to update $\theta^{(t+1)}$ and $\phi^{(t+1)}$.