Geographical Topics Learning of Geo-Tagged Social Images

Article · April 2015

DOI: 10.1109/TCYB.2015.2414489 · Source: PubMed

Xiaoming Zhang
Beihang University (BUAA)
63 PUBLICATIONS **479** CITATIONS

SEE PROFILE

Senzhang Wang
Nanjing University of Aeronautics & Astronautics
107 PUBLICATIONS **773** CITATIONS

SEE PROFILE

Geographical Topics Learning of Geo-Tagged Social Images

Xiaoming Zhang, Shufan Ji, Senzhang Wang, Zhoujun Li, *Member, IEEE*, and Xueqiang Lv

***Abstract*—With the availability of cheap location sensors, geo-tagging of images in online social media is very popular. With a large amount of geo-tagged social images, it is interesting to study how these images are shared across geographical regions and how the geographical language characteristics and vision patterns are distributed across different regions. Unlike textual document, geo-tagged social image contains multiple types of content, i.e., textual description, visual content, and geographical information. Existing approaches usually mine geographical characteristics using a subset of multiple types of image contents or combining those contents linearly, which ignore correlations between different types of contents, and their geographical distributions. Therefore, in this paper, we propose a novel method to discover geographical characteristics of geo-tagged social images using a geographical topic model called geographical topic model of social images (GTMSIs). GTMSI integrates multiple types of social image contents as well as the geographical distributions, in which image topics are modeled based on both vocabulary and visual features. In GTMSI, each region of the image would have its own topic distribution, and hence have its own language model and vision pattern. Experimental results show that our GTMSI could identify interesting topics and vision patterns, as well as provide location prediction and image tagging.**

***Index Terms*—Geographical topic, image tagging, image topic, location prediction, topic model**

I. Introduction

With rapid development of Web 2.0 and global positioning system (GPS)-equipped mobile terminals, geo-tagged social media data are tremendously increasing, which stimulates the study to identify geographical characteristics of social media sharing across different regions [18], [22]–[24]. These location-based social network services, such as Flickr, and Google Latitude, enable users to share their activities happening at certain locations in various forms. For example, the social media site Flickr provides interfaces for users to specify a location on the world map, in which the GPS records are provided together with different types of content including textual description (e.g., title, comments, and tags) and visual content. Nowadays, Flickr hosts more than 100 million geo-tagged images. It is interesting to identify how social images are popular in different regions, that is, how the language models and vision patterns of social images are distributed and correlated in different regions. For example, images about New York City might cover entirely different events compared to those about Beijing. As shown in Fig. 1, the preference of tag words and visual patterns are different for the two cities, and the correlation between tags and visual content are different for these two cities. The geographical characteristics make it possible to learn a reasonable model to identify the correlation between location distribution and image content including textual description and visual content.

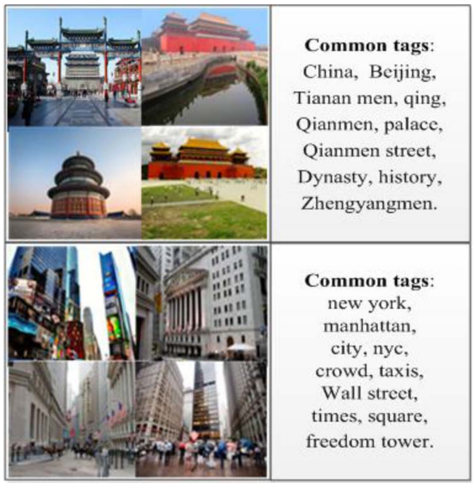


Fig. 1. Different views of Beijing and New York.

Recently, many models have been proposed to identify the relation between geographical location and image content. Unlike textual document, geo-tagged social image contains multiple types of contents, i.e., textual description, visual contents, and geographical information. Incorporating those contents simultaneously in the model to identify relations

Manuscript received August 26, 2014; revised December 9, 2014, February 16, 2015, and March 13, 2015; accepted March 15, 2015. Date of publication April 7, 2015; date of current version February 12, 2016. This work was supported in part by the National Natural Science Foundation of China under Grant 60973105, Grant 61170189, and Grant 61202239, in part by the Fundamental Research Funds for the Central Universities under Grant YWF-14-JSJXY-16, in part by the Opening Project of Beijing Key Laboratory of Internet Culture and Digital Dissemination Research under Grant ICDD201403, in part by the Innovation Foundation of Beihang Unversity for Ph.D. Graduates under Grant YWF-14-YJSY-021, and in part by the Fund of the State Key Laboratory of Software Development Environment under Grant SKLSDE-2015ZX-11. This paper was recommended by Associate Editor X. He.

X. Zhang, S. Ji, and S. Wang are with the School of Computer Science and Engineering, Beihang University, Beijing 100191, China (e-mail: yolixs@buaa.edu.cn).

Z. Li is with the School of Computer Science and Engineering, Beihang University, Beijing 100191, China, and also with the State Key Laboratory of Software Development Environment, Beihang University, Beijing 100191, China.

X. Lv is with Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing 100101, China.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2015.2414489

between image content and location is nontrivial. According to different incorporation strategies, existing studies can be classified into two categories. The first category maps images based on image contents directly, in which different types of features, e.g., visual features, and textual features, are combined linearly. Then, the location of a new image is predicted based on the mapping. For example, Hays and Efros [2] and Li *et al.* [17] proposed to model each location using a subset of visually similar images. Crandall *et al.* [3] proposed to model the geographical characteristic based on a linear combination of visual features and textual features. However, those models cannot well exploit the correlations between different types of features and the geographical distribution of the correlations. The other category of studies employ a pure language model to mine geographical language topics of social images based on textual contents and geographical information only [4], ignoring visual contents, thus ineffective to process social images with rich visual contents.

Although there exist a number of successful text-based data mining approaches or vision-based approaches of analyzing geo-tagged images, none of them addressed the following problems.

1) Geo-tagged social image contains multiple forms of content. It is very common that text, visual content, and GPS record exist simultaneously on the same social image. Incorporating this rich information may potentially help us to discover the latent information to capture the geographical characteristics of image content. However, this pursuit is nontrivial. It needs to incorporate different type of contents simultaneously using a multimodal model.
2) Visual content and textual description are correlated with each other, and the correlation is different across different regions. Thus, it is reasonable to use middle-level feature, i.e., topic, to capture the correlation between visual content and textual description and model the geographical distribution of the correlation.
3) In reality, there are also many images that are not geotagged or do not have any tags. Thus, it needs to analyze these multiple types of image contents and their correlation to support these applications, such as image location prediction and automatic image tagging.

In this paper, we propose a generative model geographical topic model of social images (GTMSIs) for geo-tagged image pattern modeling, by simultaneously incorporating geographical information, textual description, and visual contents. Geographical topic is introduced to combine geographical clustering and topic model, which model the geographical distribution of image visual features and textual words. GTMSI could identify different topic patterns across regions, where the geographical characteristics of languages and visual contents are integrated by latent topics consistently. Particularly, each region has its own topic distribution, and hence has its own language characteristics and vision patterns. A generative procedure is employed to model the production of textual and visual contents of images based on location information, and the parameters are inferred by sampling. Based on this model, image location prediction and image tagging can be implemented. Experimental results show that our GTMSI could identify interesting topics, while outperforming nontrivial baselines on location prediction and image tagging. Compared with existing works, our main contributions are as follows.

1) We propose a geographically generative model of content and locations that incorporates multiple facets of image environments in an integral fashion. Then, the "invariant" and "abstract" information of multimodal features in the geo-tagged images can be discovered.
2) A biterm treatment is used to alleviate the sparsity problem of text terms, and Bayesian treatment and sampling method is employed to infer the model parameters.
3) The proposed model outperforms several state-of-the-art models on location prediction and annotating tags for social images, as well as identifies interesting language and visual patterns in real-world social images.

The remainder of this paper is organized as follows. In the next section, we introduce related works. We briefly describe our approach in Section III and propose our geographical topic model for social image in Section IV. Section V discusses some implementation problems, and the experiments are described in Section VI. Finally, this paper is concluded in Section VII.

II. Related Works

This paper is related to image location prediction and image tagging. In the following, we review these works which are similar to this paper.

Previously, many works used the relation between image visual contents and locations to predict location directly [1]. For example, the image location is estimated by the nearest-neighbor methods, on vectors composed of image visual features [2]. Similarly, the image location is predicted based on its geo-visual neighbors [17]. There also exist feature-based geometric matching approaches, applying to co-register online famous landmark photographs for summarization [5] and browsing [6]. The geo-informative attributes are obtained for each locations based on image visual contents [7]. A soft bag-of-words method is proposed for mobile landmark recognition based on discriminative learning of image patches [39]. However, social images usually contain multiple types of contents that are correlated to geographical locations. Those works are limited to visual content analysis.

In the other side, many works use probability model to discover topics from the text content of geo-tagged images. Latent geographical topic analysis (LGTA) [4] combines geographical clustering and topic modeling to identify the geographical topics of social images, as well as estimate the topic distributions in different geographical locations for topic comparison. Another work proposes a language model based on user annotations, to place the annotated Flickr images on the map [11]. The multi-Dirichlet process (MDP)-based geographical topic model captures dependencies between geographical regions to support the detection of text topics with complex, non-Gaussian distributed spatial structures [37]. The model is based on an MDP. Those works use a pure language model to identify the geographical topic distributions of image, which do not combine visual contents in geographical topic model to identify the visual patterns of topics as well as their distributions over regions.

Similarly, some works focus on models and approaches that combine geographical modeling and language modeling to discover geographical topics from geo-tagged social media. GeoFolk uses a generative model to combine text and spatial information together, in which each topic generates latitude and longitude from two topic-specific Gaussian distributions [25]. A model based on probabilistic latent semantic indexing [26] is proposed in another work [27]. It assumes that each word is either drawn from a universal background topic or from a location and time dependent language model. A fully Bayesian generative model is also introduced to incorporate locations [8]. Then, a model [28] built upon this model is proposed, in which the notions of global topics and local topics are introduced. An even simpler approach assigns document to geodesic grids using a supervised learning method [29]. However, these models and approaches are also used to discover geographical topics from geo-tagged text documents. They cannot be applied to geo-tagged social images directly.

There are also some works combine text and visual content to mine geographical information of social images. In [9], the scale-structure identification method is adopted to extract place and event semantics for tags, based on the GPS metadata of images in Flickr. As for [3], the content analysis (based on text tags and image data) is combined with structural analysis (based on geospatial data) for image location estimation. Similarly, multiple types of contents, i.e., locations, tags, and visual features, are combined to generate diverse and representative images for the landmarks in [10].

Recently, social image tagging has also attracted great attention. In [33], tags of social images are ranked based on the importance. Hereby, the importance of the tag is determined by the number of visual neighbors of the input image which are annotated with that tag. Zhou *et al.* [15] introduced a hybrid probabilistic model for automatic image tagging. Liu *et al.* [34] proposed to rank tags associated with a given image according to their relevance to the image content. The relevance scores are estimated using a random over the tag similarity graph. Then, the top ranked tags in the visual neighbors are recommended. Zhang *et al.* [35] tagged images based on tag information capability and correlation. The information capability is a measure of the ambiguity of a tag set. Qian *et al.* [40] retagged social images with diverse semantics. However, these approaches have no consideration of the geographical distribution of image content.

III. Approach Overview

The framework of our approach is shown in Fig. 2. The input contains multiple types of content, i.e., visual content, textual description, and geographical information. The most important point of our approach is to identify the geographical language characteristics and vision patterns from geo-tagged social images, based on which location-based applications can be developed. Therefore, the core component is the social image topic model GTMSI which is proposed to model the joint distribution of both textual words and visual features, and the geographical language models and visual content patterns are reflected by the geographical topic distribution. We accomplish this thanks to the large amount of social images and the language diversity and visual content variations appearing in social images. There are many factors to influence the language and visual content used in a social image taken in a particular location. For example, textual words used in an image depend on the local culture and the visual content of the image, while the visual content depend on the local view of the geographical region, e.g., famous building, nature view, and local sports activity.

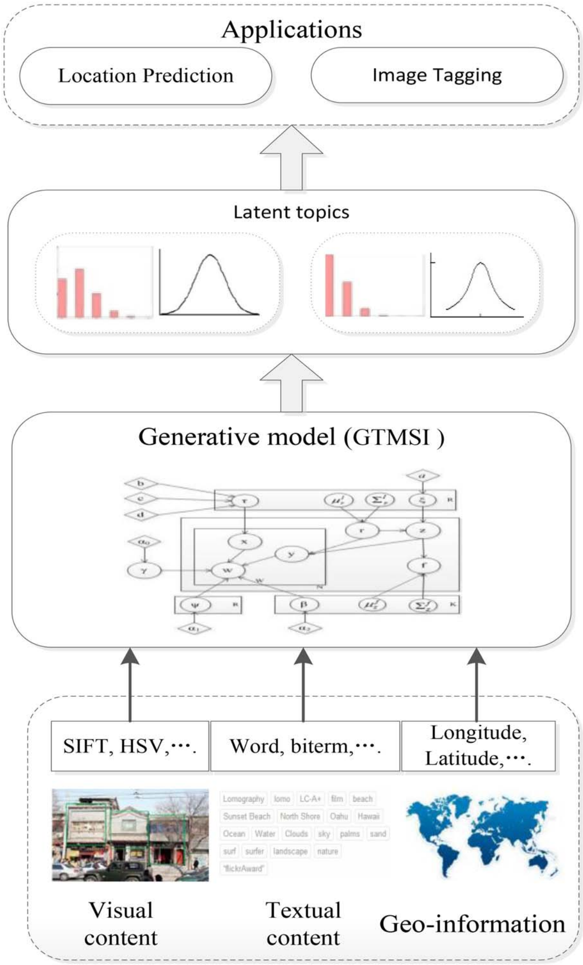


Fig. 2. Framework of our approach.

In GTMSI, each image belongs to a region, and each region has its own distribution over the topics. Here, each topic is represented by two topic-specific distributions: 1) topic-specific word distribution and 2) topic-specific distribution over visual features. The topic-specific word distribution is modeled as a multinomial distribution, and the topic-specific distribution over visual features is modeled as a normal distribution. By intertwining location and topical distribution into a joint model we are able to improve the spatial accuracy and content description. Then, the location of a new image containing only textual and visual contents is predicted based on the geographical model, i.e., estimating the joint probability of its

TABLE I
NOTATIONS USED IN THIS PAPER

Notation	Size	Description
μ^l	R^2	Mean location of a latent region
Σ^l	$R^{2\times2}$	Covariance matrix of a latent region
μ^f	$R^{\|f\|}$	Mean value of visual feature of a topic
Σ^f	$R^{\|f\|\times\|f\|}$	Covariance matrix of a latent topic
γ	$1\times W$	Background word distribution
ψ	$R\times W$	Region-specific word distribution
β	$K\times W$	Topic-specific word distribution
ξ	$R\times K$	Region-specific topic distribution
τ	$R\times3$	Region-specific topic type distribution

textual and visual contents, given a region. The mean location of the neighbor images which are most similar to the test image in the region which has the greatest joint probability is used as the predicted location of the new image. The geographical topic model can also be used to recommend tags for new image which does not contain any tags, i.e., recommending tags with the greatest conditional probabilities which are estimated by summing over the hidden variables (e.g., topic-mixing vectors and topic-specific distribution on tags, and so on).

IV. GENERATIVE MODEL

Each geo-tagged social image $p = \{w_p, f_p, l_p\}_{p\in\mathbf{P}}$ consists of three atoms: 1) w_p is a vector of words extracted from its textual contents, e.g., tags, titles, and comments; 2) f_p is the visual feature vector for the visual contents; and 3) l_p is a real-valued pair $l_p = \{l_a, l_o\}$, representing the latitude and longitude where the image is taken. For simplicity, we assume that all the textual contents in our data are generated by a fixed vocabulary of W words, and the geographical locations are clustered into R latent regions. Each topic $z \in \mathrm{Z}$ is generated from regions instead of documents. The notations are listed in Table I.

The geographical distribution of each region is assumed to be normal $N(\mu_r^l, \Sigma_r^l)_{r:1\ldots R}$, where μ_r^l and Σ_r^l are the mean vector and covariance matrix of region r, respectively. Moreover, the topic-specific visual features are also assumed to follow a normal distribution, parameterized as $(\mu^f, \sum^f) = \{(\mu_k^f, \Sigma_k^f)\}_{k:1\ldots K}$. The words that co-occur more frequently are more likely to be generated by the same topics. Similarly, the visual features that are close in feature space are more likely to appear in the same region, and they are more likely to be clustered into the some topic. Our model GTMSI has the following intuitions.

1) Words and visual features used in a social image depend on both location and topic of the image, while topics have different distributions over different regions. The topic assignments of word and visual feature are correlated with each other.
2) Different geographical regions have different language variations. Thus, a word has different probability to be used in different regions.
3) Different topics have different visual feature variations. Hence, different regions have different distributions of visual patterns.

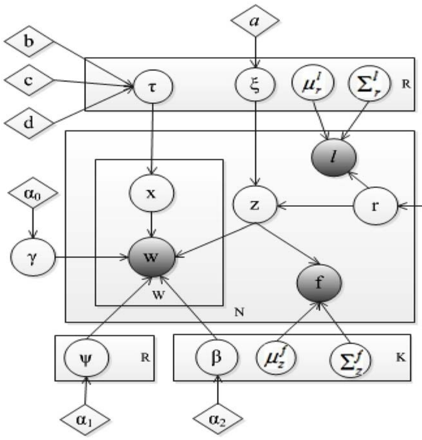


Fig. 3. Graphical representation of GTMSI.

1. Sampling a region r from the discrete distribution of region importance $\varepsilon, r \sim Discrete(\varepsilon)$.
2. Sample location l_p from normal distribution of $N(\mu_r^l, \Sigma_r^l)$.
3. To generate the visual feature f_p in image p
   i: Sample topic: $z_f^p \sim Multinomial(\xi)$.
   ii: Sample visual features: $f_p \mid z_f^p = k \sim N(\mu_k^f, \Sigma_k^f)$.
4. To generate each word $w_{p,i}$ in image p
   Sample a random variable: $x \sim Multinomial(\tau_r)$.
   If x=0
   Sample word: $w_{p,i} \sim Multinomial(\gamma)$.
   If x=1
   Sample word: $w_{p,i} \sim Multinomial(\psi_r)$.
   If x=2
   i: Sample topic: $z_w^p \sim Multinomial(\xi)$ conditioned on image features.
   ii: Sample word: $w_{p,i} \mid z_w^p = k \sim Multinomial(\beta_k)$.

Fig. 4. Generation procedure.

Fig. 3 depicts a graphical representation of GTMSI. To generate a geo-tagged image p, the generative procedure of GTMSI can be described as shown in Fig. 4. In the following sections we break the above generative steps into parts each of which addresses a specific challenge introduced by geo-tagged image.

A. Generation of Textual Words

Textual word describes the context or the semantic information of image's visual contents, which suggests that each word can be generated by correlating them to the collective set of topic indicators selected from the image generating region. However, the corpus-level words in the form of noncontent-bearing words (like: trip, 2008, and good) might appear on the top of most identified topics. As different regions have their own language characteristics, we adopt an additional machinery to handle special words and background words, which are similar to the subtraction of background and document-specific words [14]. Beside the standard latent topic produced by standard topic model [20], we introduce a background topic γ [sampled from Dir(α_0) once for the whole corpus] to generate the background words, and a region-specific topic ψ [sampled from Dir(α_1) once for each region] to generate the region-specific words. Thus, the topics in our model consist of three types, i.e., background topic, region topic, and standard latent topic which is similar to those produced by the LDA topic model [20].

The generative process for textual words now proceeds as follows. As for each image p, we associate a multinomial distribution τ on $\{0, 1, 2\}$ with prior parameters b, c, and d, which models the distributions of background words, region-specific words, and latent topic words. As shown in step 4 of Fig. 4, to generate a word w_n of image p, we first sample a random variable x from a region-specific multinomial τ_r, which in turn has Dirichlet prior parameters b, c, and d. If $x = 0$, the word is sampled from the background topic γ; and if $x = 1$, the word is sampled from the region-specific topic; else, a standard latent topic indicator, z_w^p, is selected according to topic distributions on the region and visual feature generation. There are two factors that affect the latent topic to be chosen. The first one is the occurrence of the topic in the region from which the image is sampled. The second one is the conditional probability of image visual features given the topic, i.e., the probability that the topic generates the visual feature based on the multivariate normal distribution.

B. Generation of Visual Features

The generation of visual feature is shown in step 3 of Fig. 4. Image visual features are modeled as a normal distribution whose mean and variance are topic-specific. Many works embellish the parameters of a normal distribution with an inverse Wishart prior [21], which are computationally expensive. In this paper, we take a simpler approach by placing a noninformative Jeffrey's prior over the values of the mean parameters, i.e., $\mu_z^f \sim$Unif. Meanwhile, an inverse prior over the variance is placed to penalized large variances, i.e., $P(\Sigma_z^f) \propto (\Sigma_z^f)^{-1}$ [21]. It is because that the calculation of image features might introduce noises. With such prior, the estimation of Σ_z^f for a give topic is more robust to outliers. Then, the pdf function for an image, given a topic-specific normal distribution, is revised as a function $t(f; \mu, \Sigma, n)$, which is a student t-distribution with mean μ, variance Σ, and n degree of freedom. Similarly, the pdf function $P(l_p|\mu_r^l, \Sigma_r^l)$ for a geographical location, given a region-specific normal distribution, is also revised as a student t-distribution function.

C. Inference by Gibbs Sampling

Under the generative process, we seek to compute the posterior probability

$$P\left(\gamma, \psi_{1:R}, \beta_{1:K}, \xi_{1:R}, \mu_{1:R}^l, \sum_{1:R}^{l}, \mu_{1:K}^f, \sum_{1:K}^{f} |a, b, c, d, \alpha_{0:2}, \mathbf{w}, \mathbf{f}, \mathbf{l}\right). \quad (1)$$

The above posterior probability can be easily written down from the generative model. However, the posterior is intractable. We approximate it via a collapsed Gibbs sampling procedure [12], [13], by integrating the hidden variables, e.g., the topic-mixing vectors of each region, and the topic distributions over all modalities. Therefore, the state of the sampler at each iteration contains the topic indicators for all regions. We alternate sampling each of these variables conditioned on its Markov blanket until convergence. When it converges, the expected values of all the parameters that were integrated out can be calculated. To simplify the calculation of the Gibbs sampling update equations, we keep a set of sum matrices with the form C_{xy}^{XY} to denote the number of times instance x appeared with instance y. Moreover, the subscript $-I$ is used to denote the same quantity it is added to without the contribution of item i. For example, C_{wk}^{WZ} denotes the number of times word w as sampled from latent topic k, and $C_{wk,-i}^{WZ}$ is the same as C_{wk}^{WZ} without the contribution of word w_i. The sampling procedure can be described as follows.

For each image p, a latent region r is first drawn from the following distribution, conditioned on the old topic assignments:

$$r|p, \Phi \sim P\big(r_j|\varepsilon\big)P\Big(l_p|\mu_{r_j}^l, \Sigma_{r_j}^l\Big)P\big(w_p|r_j\big)P\big(f_p|r_j\big) \quad (2)$$

where $P(l_p|\mu_{r_j}^l, \Sigma_{r_j}^l)$ is the pdf function for a multivariate normal distribution corresponding to region r_j. $P(r_j|\varepsilon)$, μ_r and Σ_r are estimated as follows:

$$\mu_{r_j}^l = \frac{1}{\mathrm{Num}(p, r_j)} \sum_{p \in \mathbf{P}} g(r(p) = r_j) l_p \quad (3)$$

$$\Sigma_{r_j}^l = \frac{\sum_{p \in \mathbf{P}} g(r(p) = r_j)\Big(l_p - \mu_{r_j}^l\Big)^T \Big(l_p - \mu_{r_j}^l\Big)}{\mathrm{Num}\big(p, r_j\big) - 1} \quad (4)$$

$$P(r_j|\varepsilon) = \frac{\sum_{p \in \mathbf{P}} g\big(r(p) = r_j\big) + \varepsilon}{|\mathbf{P}| + \varepsilon|\mathbf{R}|} \quad (5)$$

where $\mathrm{Num}(p, r_j)$ is the number of images assigned to region r_j, and $g(r(p) = r_j)$ is an indicator function which is 1 if and only if the image p is assigned to region r_j. The component $P(w_p|r_j)$ is estimated as follows:

$$\begin{aligned} P\big(w_p|r_j\big) &= \prod_{w_{p,i}} P\big(w_{p,i}|r_j\big) \\ &= \prod_{w_{p,i}} \Bigg(P\big(x = 0|r_j\big)P\big(w_{p,i}|\gamma\big) + P\big(x = 1|r_j\big)P\big(w_{p,i}|\psi_j\big) \\ &\quad + P\big(x = 2|r_j\big) \sum_{z \in Z} P\big(w_{p,i}|z\big)P\big(z|r_j\big) \Bigg) \end{aligned} \quad (6)$$

$$P\big(x = s|r_j\big) = \frac{C_{rs}^{RX} + \lambda_s(b, c, d)}{\sum_{x'} C_{rx'}^{RX} + b + c + d} \quad (7)$$

$$P\big(w_{p,i}|\gamma\big) = \frac{C_{i0}^{W0} + \alpha_0}{\sum_{w'} C_{w'0}^{W0} + W\alpha_0} \quad (8)$$

$$P\big(w_{p,i}|\psi_j\big) = \frac{C_{ir}^{WR} + \alpha_1}{\sum_{w'} C_{w'r}^{WR} + W\alpha_1} \quad (9)$$

$$P\big(w_{p,i}|z = k\big) = \frac{C_{ik}^{WZ} + \alpha_2}{\sum_{w'} C_{w'k}^{WZ} + W\alpha_2} \quad (10)$$

$$P\big(z=k|r_j\big)=\frac{\big(C_{rk}^{RZ}+\alpha_1\big)*p_{f_pk}^{FZ}}{\sum_{k'}\Big(\Big(C_{rk'}^{RZ}+\alpha_1\Big)*p_{f_pk'}^{FZ}\Big)}\tag{11}$$

where $\lambda_s(b,c,d)=b,c,$ and d for $s=0$, 1, and 2, respectively, and the variable x act as a switch: if $x=2$, the word is generated by the standard topic production mechanism, whereas if $x=0$ or $x=1$, the word is sampled from a background specific multinomial or a region-specific multinomial. C_{rs}^{RX} counts the number of times region r is assigned the background topic ($s=0$), the number of times region r is assigned the region-specific topic ($s=1$), and the number of times region r is assigned a standard latent topic ($s=2$). $p_{f_pk}^{FZ}$ is the probability that topic k generates the visual feature f_p based on the multivariate normal distribution. $P(f_p|r_j)$ is estimated as follows:

$$P\big(f_p|r_j\big)=\sum_{k=1}^{K}P\big(f_p|z=k\big)P\big(\mathrm{z}=k|r_j\big)\tag{12}$$

where $P(f_p|z=k)=P(f_p|\mu_k^f,\Sigma_k^f)$ is the productive probability of the visual features, the pdf of a multivariate normal distribution, μ_k^f is the sample mean of the values of image feature that are assigned to topic k, and Σ_k^f is defined similarly.

Then, we update the topic assignments for the textual words and visual features of image p conditioned on the region assignment as follows.

Sample a topic for each word token $w_{p,i}$ in p

$$\begin{aligned}&P(x_i=0|w,x_{-i},z_{-i},\alpha_0,b,c,d)\\&\quad\propto\frac{C_{r0,-i}^{RX}+b}{\sum_{x'}C_{rx',-i}^{RX}+b+c+d}\cdot\frac{C_{w,-i}^{W0}+\alpha_0}{\sum_{w'}C_{w',-i}^{W0}+W\alpha_0}\end{aligned}\tag{13}$$

$$\begin{aligned}&P(x_i=1|w,x_{-i},z_{-i},\alpha_1,b,c,d)\\&\quad\propto\frac{C_{r1,-i}^{RX}+c}{\sum_{x'}C_{rx',-i}^{RX}+b+c+d}\cdot\frac{C_{wr,-i}^{WR}+\alpha_1}{\sum_{w'}C_{w'r,-i}^{WR}+W\alpha_1}\end{aligned}\tag{14}$$

$$\begin{aligned}&P(x_i=2,z=k|w,x_{-i},z_{-i},a,\alpha_2,b,c,d)\\&\quad\propto\frac{C_{r2,-i}^{RX}+d}{\sum_{x'}C_{rx',-i}^{XR}+b+c+d}\cdot\frac{\Big(C_{rk,-i}^{RZ}+a\Big)\cdot p_{f_pk}^{FZ}}{\sum_{k'}\Big(\Big(C_{rk',-i}^{RZ}+a\Big)\cdot p_{f_pk'}^{ZF}\Big)}\\&\qquad\cdot\frac{C_{wk,-i}^{WZ}+\alpha_2}{\sum_{w'}C_{w'k,-i}^{WZ}+W\alpha_2}.\end{aligned}\tag{15}$$

Equation (15) indicates that the topic assignment for a word is affected by the region preference of the topic and the topic preference of the corresponding visual features.

Finally, sample a topic for the visual feature f_p

$$\begin{aligned}P\big(z=k|f_p,w_p\big)\propto{}&\frac{C_{rk,-p}^{RZ}+a}{\sum_{k'}\Big(C_{rk',-p}^{RZ}+a\Big)}\cdot P\big(f_p|z=k\big)\\&\cdot\ \mathrm{Unif}\big(w_p|z=k\big)\end{aligned}\tag{16}$$

where w_p is the word vector of image p, and the first part measures the comparability of joining a topic, given the region. As described above, the pdf function for an image, given a topic-specific normal distribution, can be revised as a function $t(f;\mu,\Sigma,n)$ of a student t-distribution. Thus, the second part is calculated by

$$P\big(f_p|z=k\big)\propto t\Big(f;\hat{\mu}_k^f,\hat{\Sigma}_k^f,C_{f_pk,-p}^{FZ}-1\Big)\tag{17}$$

where $\hat{\mu}_k^f$ is the sample mean of the values of image features assigned to topic k, $\hat{\Sigma}_k^f$ is defined similarly, and $C_{f_pk}^{FZ}$ is the number of times that image features are sampled from topic k. Additionally, the pdf function $P(l_p|\mu_{r_j}^l,\Sigma_{r_j}^l)$ in (2) can also be calculated similarly. Since other textual words in the image are generated conditioned on the topic indicator of the visual features, the topic assignment should take those textual words into consideration. Therefore, the third part is used to measure how likely the current assignment on the topic indicators of other words is given the new assignment to this image feature's topic indicator, calculated as follows:

$$\mathrm{Unif}\big(w_p|z=k\big)=\prod_{i=1}^{N_{w_p}}\frac{C_{rz_i,-p}^{RZ}+g(z_i=k)}{\sum_{k'}C_{rk',-p}^{RZ}+g(z_i=k)}\tag{18}$$

where N_{w_p} is the number of words in image p, and $g(z_i=k)$ is an indicator function, equal to 1 if and only if the express inside is evaluated to be true.

V. Implementation Notes

Usually, topic model of social media document suffer from the severe data sparsity in short text content [31]. In Flickr, the text content of many geo-tagged images contains only several tags. We alleviate the effect of data sparsity problem by enriching textual terms. Beside individual words and tags, the biterms (i.e., word co-occurrence patterns) are also used as the textual features of image. A biterm denotes an unordered word-pair co-occurring in an image's text content. For example, in an image' text content, i.e., tags ("NY," "Manhattan," and "WTC"), there are three biterms, i.e., "NY Manhattan," "NY WTC," and "Manhattan WTC." Thus, topic-specific distribution on words in the model is extended to also include distribution on pairs of correlated words, and the vocabulary is also extended to contain the extracted biterms. For example, if the tags NY and Manhattan frequently co-occur with each other in the same images, we can identify that the biterm NY Manhattan and these individual tags belong to the same topic (i.e., New York City). In some cases, biterm is more discriminative than an individual tag with single word. For example, "times–plaza" denote a specific location of New York City, while the meaning of times or plaza is too broad.

To simplify the implementation, we consider that each biterm is drawn from a specific topic independently. The probability that a biterm drawn from a specific topic is further captured by the chances that both words in the biterm are drawn from the topic. Thus, the joint probability of generating a biterm $t=(w_i,w_j)$ can be denoted by

$$\begin{aligned}P(t|r)={}&P(x=0|r)P(w_i|\gamma)P(w_j|\gamma)\\&+P(x=1|r)P(w_i|\psi)(w_j|\psi)\\&+P(x=2|r)\sum_{z\in Z}P(w_i|z)P(w_j|z)P(z|r_j)).\end{aligned}\tag{19}$$

```
Algorithm 1: Gibbs sampling algorithm for GTMSI
Input: the number of topics, hyperparameters, training image set.
Output: multinomial parameter $\xi$, $\tau$, $\psi$, $\beta$.
Build the biterm set;
initialize region topic assignments randomly for all the terms
for $iter = 1$ to $N_{iter}$ do
    for each image $p \in P$ do
        draw region indicator $r$ by using (2);
        for each term in $p$
            draw topic indicator for each term by using (13-15), or similar equations as (20);
        end for term;
        draw topic indicator for visual feature $f_p$ by using (16);
        update count matrixes $C_{xy}^{XY}$ and parameters of the normal distributions;
    end for image;
end for iterations;
```

Fig. 5. Procedure of Gibbs sampling.

TABLE II
EXAMPLES OF GEOGRAPHICAL LANGUAGE MODELS

location	Top ranked terms
New York City	New York, NY city, empire state building, manhattan, Freedom Tower, central park, museum, Brooklyn.
Chicago	Chicago, lake Michigan, Jackson Park, Willis Tower, Illinois, Chicago River, Michigan
San Francisco	San Francisco, bay area, bay bridge, golden gate, island, cable car, Transamerica pyramid
Philadelphia	Philadelphia, Independance Hall, baseball, eagles, philly, Logan Square, Museum
Miami	Florida, Biscayne Bay, Everglades, Caribbean sea, Caribbean island, palmTree, downtown Miami, bayside

Then, the inferring algorithm is also revised to sample the topic indicator of biterm. For example, to sample the latent topic of a biterm $t = (w_i, w_j)$, we can obtain the conditional probability by applying the chain rule on the joint probability of the two words [32]

$$
\begin{aligned}
&P(x_t = 2, z = k | w, x_{-t}, z_{-t}, a, \alpha_2, b, c, d)\\
&\quad\propto \frac{C^{RX}_{r2,-t} + d}{\sum_{x'} C^{XR}_{rx',-t} + b + c + d} \cdot \frac{\left(C^{RZ}_{rk,-t} + a\right) \cdot p^{FZ}_{f_p k}}{\sum_{k'} \left(\left(C^{RZ}_{rk',-t} + a\right) \cdot p^{ZF}_{f_p k'} \right)}\\
&\qquad \cdot \frac{\left(C^{WZ}_{wk,-i} + \alpha_2\right)\left(C^{WZ}_{wk,-j} + \alpha_2\right)}{\left(\sum_{w'} C^{WZ}_{w'k,-i} + W\alpha_2\right)^2}. \qquad (20)
\end{aligned}
$$

An overview of the Gibbs sampling procedure is shown in Fig. 5. Due to space limitation, we omit the detailed derivation of it. The major time consuming part in the Gibbs sampling procedure of GTMSI is drawing the topic assignment for every term (including originally textual word tokens and biterms) in images, with time complexity $O(K|P|l')$, where $l' = \sum_i n_{p_i}/|P|$ is the average number of terms per image in the training set. The total time complexity is about $O(N_{\text{iter}}(|P| + K|P|l'))$.

VI. EXPERIMENTS

In this section, we study the effectiveness of GTMSI on two real-world datasets.

A. Dataset

The first dataset is a set of photos released by the MediaEval2012 placing task [36]. The set contains 3 185 258 geo-tagged Flickr (http://www.flickr.com/) photos randomly sampled with a method that attempts to maintain coverage of the globe. Since this release dataset includes only the metadata and not the images themselves, we download the raw images using the links in the metadata and randomly select one million images to use as a dataset for our experiments. The scale invariant feature transform (SIFT) feature [30] is used to represent the visual content of each image. Specifically, we sample a fixed number of keypoints per image, and all the SIFT features of these keypoints are clustered to create a "visual vocabulary." For the dataset MediaEval2012, we set the number of clusters to be 1000. Then, each image is represented by a 1000-dimensional vector indicating how many times each SIFT "keyword" occurs in the image.

In addition, we evaluated the location prediction performance with a group of geo-tagged images taken from NUS-WIDE dataset [19]. NUS-WIDE database contains about 50 thousands geo-tagged images, and the 500-D bag of visual words is selected as visual features of this dataset. We divided the two datasets into 80% for training and 20% for testing.

For those two datasets, the text content denotes image's tags, title, comments, and textual words are extracted from image's text content. However, some tags contain more than one word. To avoid destroying the semantic information of these tags, we take the whole tag as a single word token. Then, we remove the tags which contain more than one word from the textual content and combine any two distinct words in the remaining textual content of an image to form a biterm.

B. Topic Learning Results

Here, we present a few topics identified by GTMSI from the MediaEval2012 dataset, in which the number of topic is fixed to 80. Table II shows four examples of geographical language models. In GTMSI, regions are variables, representing no real-world cities or districts. Thus, we assign the mean vectors of regions with the nearest cities for description. It turns out that most top ranked words are actually the name of locations and famous buildings. In addition, we find that top ranked terms in different regions vary significantly. Thus, the language

TABLE III
EXAMPLES OF LATENT TOPICS

Topic	Top ranked terms	Top ranked images
city	Cityscape, skyline, city light, Street, car, building, architecture, skyscraper towers.	
zoo	Zoo, park, animal, great, nature, lion, tree monkeys, black pool, tiger.	
river	river, boat, bridge, water, sky, reflect, windy, blue, bank, road.	
beach	beach, ocean, pacific ocean, people, sea, girl, walk, sand beach	

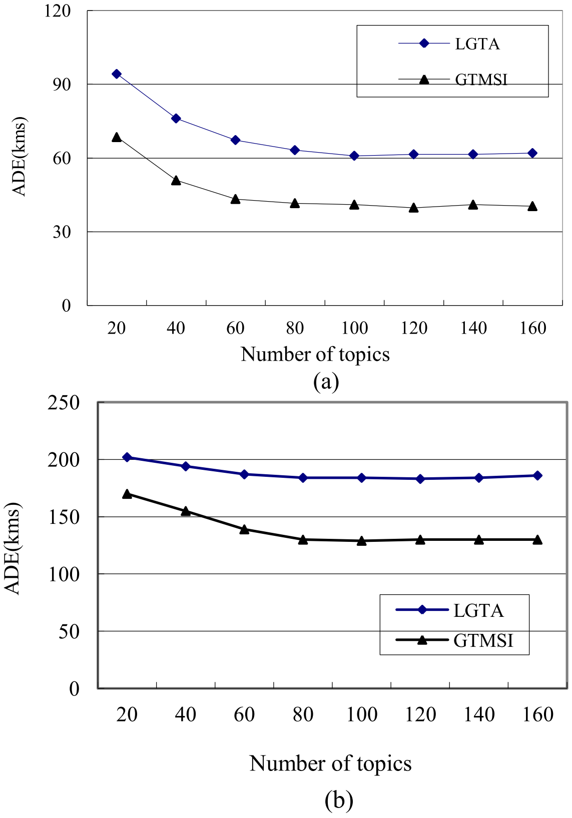
models are effective to distinguish different regions and predict locations. Table III shows some examples of the standard latent topics, which are manually selected and assigned titles. Here, the models are used to capture broader textual topics and visual patterns. Compared to regional language models, those language models are relatively broader, across different regions.

C. Image Location Prediction

Here, we are to study the performance of GTMSI on image location prediction. Other four approaches are used as baselines in this set of experiments.

The first one is LGTA [4] which utilizes the topic model only on the tags of image. It clusters images into regions according to the geographical information. Each region has a multinomial distribution on the topic which also has a multinomial distribution on the words. We realize the prediction of this model by two steps. First, the region index which maximizes the test image likelihood is selected. Then, the mean location of a set of most similar images of the region is used as the predicted location.

The second one is IM2GPS [2] which use only visual content. It uses the visual distance to find the 130 nearest neighbors and derive the location from these geo-tagged neighbors. To derive the location of the test image, mean-shift [16] (scale = 0.00001) is used to cluster the neighbors based on the location information. Then, the mean location of the cluster which has the highest cardinality is used as the location of the test image.



(a)

(b)

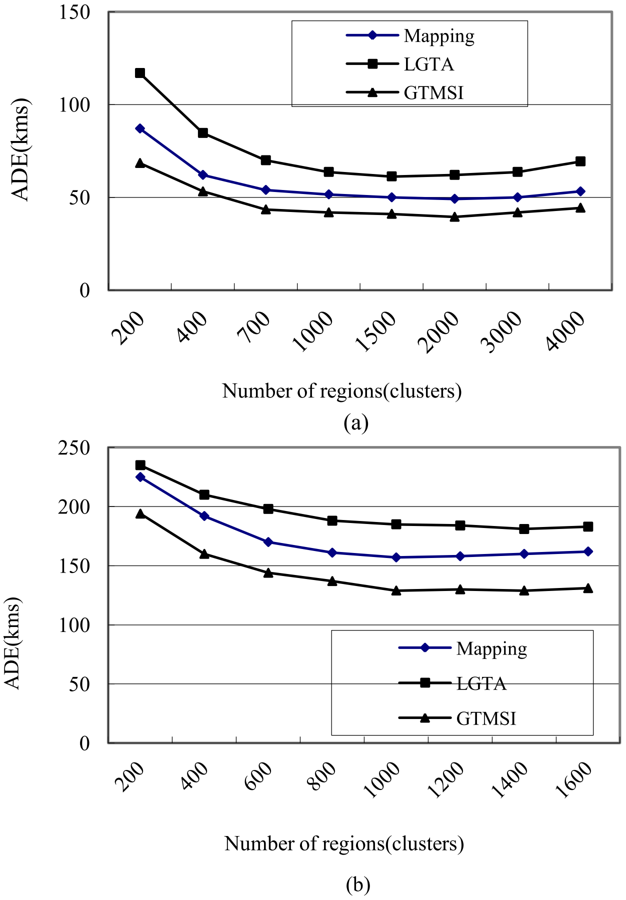
Fig. 6. Model comparison with different number of topics. (a) MediaEval2012 and (b) NUS-WIDE datasets.

The third one is GVR [17] which uses only visual content. It first retrieves a set of top-K visual-neighbors as candidates. Then the candidate whose geographical neighbors that are also in the candidates set are most visually similar to the test image is selected as the target, and the location of the target image is used as the location of the test image.

The fourth baseline approach called mapping [3] which uses both vision and text. Since the two datasets does not contain temporal and personal information, we use the first method proposed by this author. That is, it first cluster images based on the geographical information. Then the test image is classified into a cluster based on a linear combination of visual feature and textual tags, and the mean location of this cluster is used as the location of the test image.

As for our approach, the prediction procedure contains two steps. The region that maximizes the likelihood of the test image's textual and visual content is selected. Then, the mean location of the images that are most similar to the test images in the selected region is used as the location of the test image. For performance evaluation, we calculate the Euclidean distance between the predicted location and real location, using the metric of average distance error (ADE) calculated as follows:

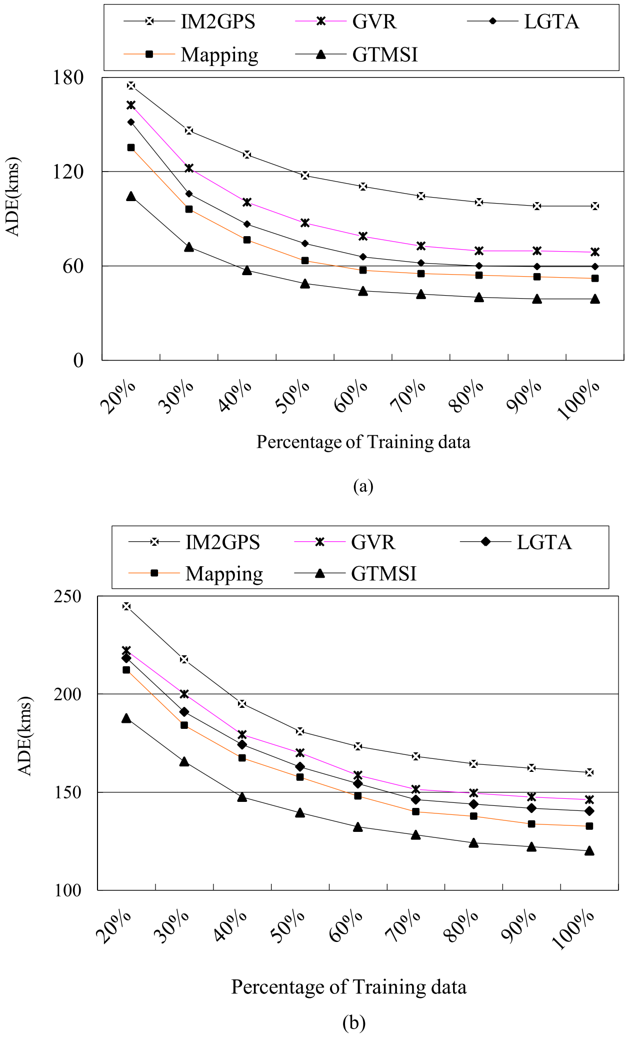
$$\mathrm{ADE} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{dis}\left(\hat{l}_i, l_i\right) \tag{21}$$



(a)

(b)

Fig. 7. Model comparison with different number of regions. (a) MediaEval2012 and (b) NUS-WIDE datasets.



(a)

(b)

Fig. 8. Model comparison with different percentages of training data. (a) MediaEval2012 and (b) NUS-WIDE datasets.

where N is the total number of test images, and $\mathrm{dis}(\hat{l}_i, l_i)$ is the Euclidean distance between the predicted location and true location.

As LGTA has the topic number parameter, we compare GTMSI against LGTA with different numbers of topics, with a fixed region number 1500. As shown in Fig. 6, the average distance does not change greatly as the number of topics varies, which might because that, those models predict the location based on the mean vectors of latent regions. Fixing the number of regions is approximate to fixing the range of regions. Thus, a fixed number of regions would bound the prediction performance to some range. However, it is clear that GTMSI significantly outperforms LGTA. It is because that GTMSI learns special words for different regions, which are helpful to discriminate different regions. Moreover, the visual features as well as their relations with textual contents are exploited in GTMSI, which is complementary to geographical language model mining. As different region has its own vision patterns, the language model is inadequate to discriminate the geographical characteristic of different regions, which might explain why LGTA performs worse.

Fig. 7 shows the comparison of different approaches with various numbers of region and the number of topic fixed to 100. It is clear that GTMSI outperforms other approaches significantly. As the number of regions increases, the performances of those approaches improve significantly at the early stage, and then deliver no obvious change. It is because that the prediction is based on the mean locations of latent regions. Thus, the more regions the model takes, the more flexible the prediction would be. It is relatively difficult to predict the real region when large amounts of regions are taken, resulting in image misallocation, thus degrading the performance. It can be concluded that Mapping is more effective than LGTA. This is because that Mapping exploits both visual features and textual content to predict location, while LGTA mainly depend on the textual content of image. However, the approach mapping combines different types of social image content using a linear method, and it classify image into different regions based on a linear combination of textual vector and visual features. The inherent correlation between different types of content cannot be effectively exploited to model the geographical characteristics. Since image content are of wide diversity, many regions may overlap in the feature space. The method of direct classifying may locate the image in a wrong region.

Then, we evaluate the performance of all the approaches by using different percentages of training data. Fig. 8 shows the experimental results with the percentage of training dataset varied from 20% to 100%. It can be concluded that the all the performances are affected by the volume of training data. This is might because that when the training dataset is too small, the distributions of images in many locations are very sparse, especially in the locations where are less frequently photographed. Therefore, the prediction of these locations is less effective. It is interesting to find that text information is help in location prediction, since LGTA and mapping outperform other two vision-based approaches. The performances of both IM2GPS and GVR have a high dependency on the selection of visually similar images. However, the visually similar images might be semantically dissimilar images due to the "semantic gap" problem [38]. Thus, visually similar images could be far away from each other. Our model integrates different types of contents, using their correlations to identify the latent relation between locations and image's textual and visual contents, which is more effective in location prediction.

D. Image Tagging

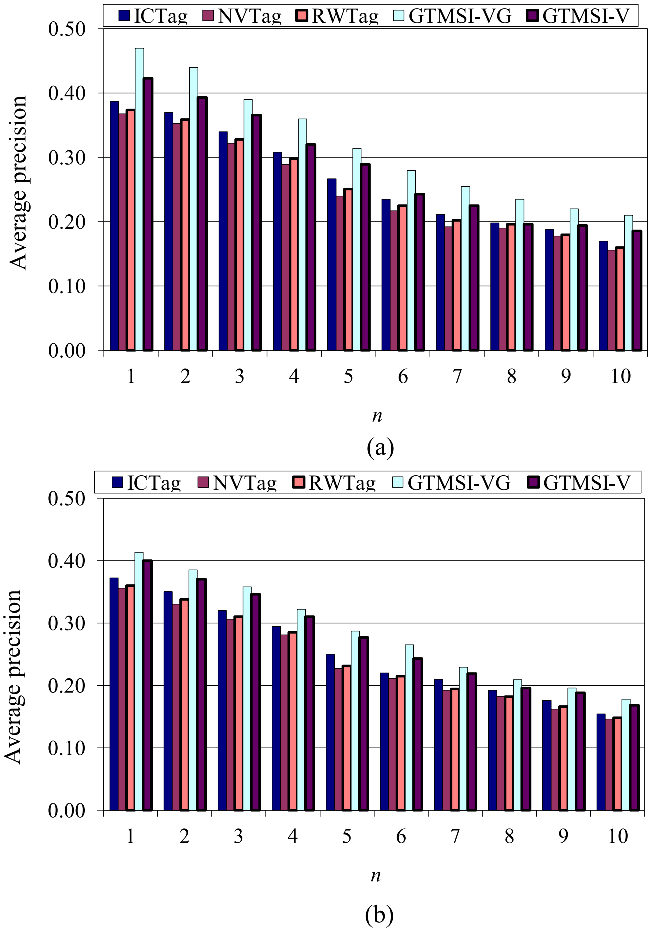
In this set of experiments, we test the performance of our topic model used for the task of image tagging. Image tagging is to automatically assign image with semantic words called tags [33]–[35]. We take the image tagging as a tag ranking problem, and the top ranked terms are used as the tags of the test image. As for the content of test image, there are two cases, i.e., image with visual content and geographical information, and image with only visual content. In the first case, the assigned tags depend on both of the visual content and geographical information. Given an image with visual feature f and location l, the tags are ranked based on the posterior probability estimated as follows:

$$\begin{aligned} P(t|f,l) &\propto \sum_{r} P(t|f,r)P(r|l) \\ &\propto \sum_{r} P(t|f,r)P\Big(l|\mu_r^l,\Sigma_r^l\Big)P(r|\varepsilon) \\ &\propto \sum_{r}\Bigg(P(x=0|r)P(t|\gamma)+P(x=1|r)P(t|\psi_r) \\ &\qquad + P(x=2|r_j)\sum_{z\in Z}P(t|z)P\Big(f|\mu_z^f,\Sigma_z^f\Big)P(z|r)\Bigg). \end{aligned} \tag{22}$$

For implementation simplicity, in the second case, we assume that the assigned tags mainly depend on the visual content and its topic distribution. Therefore, given an image with visual feature f, the tags are ranked based on the posterior probability simply estimated as follows:

$$\begin{aligned} P(t|f) &\propto \sum_{z} P(t|z)P(z|f) \\ &\propto \sum_{z\in Z} P(t|z)P\Big(f\Big|\mu_z^f,\Sigma_z^f\Big). \end{aligned} \tag{23}$$

Three baseline approaches, i.e., neighbor voting based image tagging (NVTag) [33], tagging based on tag information capability and correlation (ICTag) [35], and random walk based image tagging (RWTag) [34], are evaluated for comparison. These baseline approaches annotate image only based on visual similarity, in which the geographical information is not exploited. A traditional performance evaluation criterion, i.e., the precision at top-n annotated tag, is used to evaluate the performance of image tagging.



(a)

(b)

Fig. 9. Average precision at top-n annotated tags. (a) MediaEval2012 and (b) NUS-WIDE datasets.

Fig. 9 shows the average precision at top-n annotated tags with n varied from 1 to 10. Table IV gives some examples of image tagging. GTMSI-VG denotes the first case that test image contains both of visual content and geographical information, and GTMSI-V denotes the second case that test image contains only visual content. From these figures, we can see that our approach archives encouraging improvement on this measure, in which the average precision is greatly improved. Several conclusions can be drawn from the figures. First, on the large-scale real-world image dataset, our proposed approach outperforms other approaches significantly on different number of annotated tags per test image. The improvement is supposed to stem from the fact that GTMSI-VG use the location information and both of our approaches exploit the latent correlation between different types of content. However, NVTag mainly depend on the visual similarity. RWTag rank the tags associated with a given image based on visual similarity and tag co-occurrence, and the annotated tags are top tags of each neighboring image according to its tag ranking list. Thus, many common tags in these visual neighbor images are selected. Due to the semantic problem, these selected tags

TABLE IV
EXAMPLES OF IMAGE TAGGING

Images			
Ground truth	Chicago, Illinois, Midwest, people, cloudgate, bean, chicagoist, Millennium Park	London, BigBen, Westminster, bridge, lights, clouds, night, water	Jinshanling, Beijing, China, greate wall, tree, mountain, clear
GTMSI-VG	Chicago, Illinois, cloudgate, clouds, building, people,	London, BigBen, water, bridge, boat, clouds	Beijing, China, great wall, tree, mountain, cloud,
ICTag	Chicago, Cloudgate, searstower, Hancock, building, car,	London, BigBen, londoneye, Towerbridge, Street, tatemodern	Beijing, China, great wall, Tiananmen, badaling, Jinshanling,
RWTag	People, building, Street, car, skyscraper, tree	ocean, BigBen, bridge, lights, tourism, water	tree, clouds, mist, great wall, hill, gate,
NVTag	cloudgate, People, car, skyscraper, clouds, building	BigBen, London, water, boat, people, sky	building, clouds, hill, sunset, tree, great wall,

may not be related to the visual content of the query image. Our approach mine the geographical correlation between tags and visual content, by which the tags which reflect the image's visual content more effectively can be selected to annotate the query image. ICTag also uses a relatively loose way to combine tag co-occurrence and visual similarity to recommend tags. Second, it shows that GTMSI-VG outperform GTMSI-V and other approaches, which means that image's geographical information is helpful to recommend appropriate tags for this image. This is because that each geographical region has its own language pattern and tagging characteristic as shown in Fig. 1. Moreover, a true geographic coordinates is more effective in locate image's geographical region. Therefore, GTMSI-VG outperforms GTMSI-V which ranks tag based on geographically latent topic distribution only. Third, with the n increase, the precision of all approaches decease. This is because that the probability of including noisy tags increases with the number of result tags increase, and also many manually labeled tags are noisy. However, the proposed algorithm keeps higher average precision than other approaches over all numbers of annotated tags.

VII. CONCLUSION

The emerging trend of geo-tagged social image stimulates a wide variety of novel researches and applications. In this paper, we address the problem of mining geographical topics of geo-tagged social images by introducing a GTMSI, which simultaneously incorporates multiple types of image contents, i.e., textual description, visual contents, and location information. GTMSI could identify both language models and vision patterns across different geographical regions. Two real-life datasets and several baselines are used for comparative studies. Experimental results show that GTMSI is effective in interesting topic identification as well as location prediction and tagging, for new images. GTMSI would be widely applied to geo-based clustering, geo-based image retrieval, geographical event detection, point of interest recommendation, etc., which could also be further extended to process social audios and videos.

As for future works, we would introduce other essential information into GTMSI, such as users' social connection, relation between images and users, temporary information, human route, and so on.

REFERENCES

[1] W. B. Thompson, C. M. Valiquette, B. H. Bennett, and K. T. Sutherland, "Geometric reasoning under uncertainty for map-based localization," *Spat. Cogn. Comput.*, vol. 1, no. 3, pp. 291–321, 1999.

[2] J. Hays and A. Efros, "IM2GPS: Estimating geographic information from a single image," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Anchorage, AK, USA, 2008, pp. 1–8.

[3] D. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg, "Mapping the world's photos," in *Proc. 18th Int. Conf. World Wide Web (WWW)*, Madrid, Spain, 2009, pp. 761–770.

[4] Z. Yin, L. Cao, J. Han, C. Zhai, and T. Huang, "Geographical topic discovery and comparison," in *Proc. 20th Int. Conf. World Wide Web (WWW)*, Hyderabad, India, 2011, pp. 247–256.

[5] I. Simon, N. Snavely, and S. Seitz, "Scene summarization for online image collections," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Rio de Janeiro, Brazil, 2007, pp. 1–8.

[6] N. Snavely, S. Seitz, and R. Szeliski, "Photo tourism: Exploring photo collections in 3D," in *Proc. Spec. Interest Group Comput. Graph. Interact. Tech. Conf. (SIGGRAPH)*, Boston, MA, USA, 2006, pp. 835–846.

[7] Q. Fang, J. Sang, and C. Xu, "GIANT: Geo-informative attributes for location recognition and exploration," in *Proc. ACM Int. Conf. Multimedia (MM)*, Barcelona, Spain, 2013, pp. 13–22.

[8] C. Wang, J. Wang, X. Xie, and W. Y. Ma, "Mining geographic knowledge using location aware topic model," in *Proc. ACM Workshop Geograph. Inf. Retrieval*, Lisbon, Portugal, 2007, pp. 65–70.

[9] T. Rattenbury, N. Good, and M. Naaman, "Towards automatic extraction of event and place semantics from Flickr tags," in *Proc. 30th Annu. Int. Conf. Res. Develop. Inf. Retrieval (SIGIR)*, Amsterdam, The Netherlands, 2007, pp. 103–110.

[10] L. S. Kennedy and M. Naaman, "Generating diverse and representative image search results for landmarks," in *Proc. 17th Int. Conf. World Wide Web (WWW)*, Beijing, China, 2008, pp. 297–306.

[11] P. Serdyukov, V. Murdock, and R. V. Zwol, "Placing Flickr photos on a map," in *Proc. 32nd Int. Conf. Res. Develop. Inf. Retrieval (SIGIR)*, Boston, MA, USA, 2009, pp. 484–491.

[12] H. M. Wallach, "Topic modeling: Beyond bag-of-words," in *Proc. Int. Conf. Machine. Learn. (ICML)*, Pittsburgh, PA, USA, 2006, pp. 977–984.

[13] T. L. Griffiths and M. Steyvers, "Finding scientific topics," in *Proc. Nat. Acad. Sci.*, Washington, DC, USA, 2004, pp. 5228–5235.

[14] C. Chemudugunta, P. Smyth, and M. Steyvers, "Modeling general and specific aspects of documents with a probabilistic topic model," in *Proc. Neural Inf. Process. Syst. (NIPS)*, Vancouver, BC, Canada, 2006, pp. 241–248.

[15] N. Zhou, W. K. Cheung, G. Qiu, and X. Xue, "A hybrid probabilistic model for unified collaborative and content-based image tagging," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 7, pp. 1281–1294, Jul. 2011.

[16] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, May 2002.

[17] X. C. Li, M. Larson, and A. Hanjalic, "Geo-visual ranking for location prediction of social images," in *Proc. 3rd ACM Conf. Int. Conf. Multimedia Retrieval (ICMR)*, Dallas, TX, USA, 2013, pp. 81–88.

[18] Y. Chen *et al.*, "From interest to function: Location estimation in social media," in *Proc. 27th AAAI Conf. Artif. Intell. (AAAI)*, Bellevue, WA, USA, 2013, pp. 180–186.

[19] T.-S. Chua *et al.*, "NUS-WIDE: A real-world web image database from National University of Singapore," in *Proc. Int. Conf. Image Video Retrieval (CIVR)*, Santorini, Greece, 2009, pp. 1–9.

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

[21] M. J. Daniels, "A prior for the variance in hierarchical models," *Can. J. Stat.*, vol. 27, no. 3, pp. 567–578, 1999.

[22] Z. Song, M. Ester, and B. Bhattacharya, "Discovering more meaningful regions: A regularized geographical topic model," in *Proc. 22nd Int. Conf. World Wide Web (WWW)*, Rio de Janeiro, Brazil, 2013, pp. 231–240.

[23] L. Hong, A. Ahmed, S. Gurumurthy, A. Smola, and K. Tsioutsiouliklis, "Discovering geographical topics in the twitter stream," in *Proc. 21st Int. Conf. World Wide Web (WWW)*, Lyon, France, 2012, pp. 769–778.

[24] A. Ahmed, L. Hong, and A. Smola, "Hierarchical geographical modeling of user locations from social media posts," in *Proc. 22nd Int. Conf. World Wide Web (WWW)*, Rio de Janeiro, Brazil, 2013, pp. 25–36.

[25] S. Sizov, "GeoFolk: Latent spatial semantics in Web 2.0 social media," in *Proc. 3rd Int. Conf. Web Search Data Min. (WSDM)*, New York, NY, USA, 2010, pp. 281–290.

[26] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Mach. Learn.*, vol. 42, no. 2, pp. 177–196, 2001.

[27] Q. Mei, C. Liu, H. Su, and C. Zhai, "A probabilistic approach to spatiotemporal theme pattern mining on weblogs," in *Proc. 15th Int. Conf. World Wide Web (WWW)*, Southampton, U.K., 2006, pp. 533–542.

[28] Q. Hao *et al.*, "Equip tourists with knowledge mined from travelogues," in *Proc. 19th Int. Conf. World Wide Web (WWW)*, Raleigh, NC, USA, 2010, pp. 401–410.

[29] B. P. Wing and J. Baldridge, "Simple supervised document geolocation with geodesic grids," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguist. Human Lang. Technol. (ACL)*, Portland, OR, USA, 2011, pp. 955–964.

[30] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[31] L. Hong and B. Davison, "Empirical study of topic modeling in twitter," in *Proc. 1st Workshop Soc. Media Anal.*, Washington, DC, USA, 2010, pp. 80–88.

[32] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A biterm topic model for short texts," in *Proc. 22nd Int. Conf. World Wide Web (WWW)*, Rio de Janeiro, Brazil, 2013, pp. 1445–1456.

[33] X. Li, C. G. Snoek, and M. Worring, "Learning social tag relevance by neighbor voting," *IEEE Trans. Multimedia*, vol. 11, no. 7, pp. 1310–1322, Nov. 2009.

[34] D. Liu, M. Wang, X. S. Hua, and H. J. Zhang, "Tag ranking," in *Proc. 18th Int. Conf. World Wide Web (WWW)*, Madrid, Spain, 2009, pp. 351–360.

[35] X. Zhang, H. T. Shen, Z. Huang, Z. Li, and Y. Yang, "Automatic tagging by exploring tag information capability and correlation," *World Wide Web*, vol. 15, no. 3, pp. 233–256, 2012.

[36] A. Rae and P. Kelm, "Working notes for the placing task at MediaEval 2012," in *Proc. MediaEval Workshop*, Pisa, Italy, 2012, pp. 1–2.

[37] C. C. Kling, J. Kunegis, S. Sizov, and S. Staab, "Detecting non-Gaussian geographical topics in tagged photo collections," in *Proc. Int. Conf. Web Search Data Min. (WSDM)*, New York, NY, USA, 2014, pp. 603–612.

[38] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, Dec. 2000.

[39] T. Chen and K.-H. Yap, "Discriminative BoW framework for mobile landmark recognition," *IEEE Trans. Cybern.*, vol. 44, no. 5, pp. 695–706, May 2014.

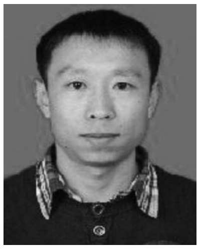
[40] X. Qian, X.-S. Hua, Y. Y. Tang, and T. Mei, "Social image tagging with diverse semantics," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2493–2508, Dec. 2014.

Xiaoming Zhang received the B.Sc. and M.Sc. degrees from the National University of Defence Technology, Changsha, China, in 2003 and 2007, respectively, both in computer science and technology. He received the Ph.D. degree in computer science from Beihang University, Beijing, China, in 2007.

Since 2007, he has been the Lecturer with Beihang University. His current research interests include social media analysis, image tagging, and topic detection and tracking.

Shufan Ji received the M.Sc. degree from Nanjing University, Nanjing, China, and the Ph.D. degree from National University of Singapore, Singapore, both in computer science.

From 2007 to 2008, she was a Post-Doctoral Fellow with Stanford University, Stanford, CA, USA. She is currently with the School of Computer Science and Engineering, Beihang University, Beijing, China, where she has been the Associate Professor, since 2012. Her current research interests include data mining and bioinformatics science. She has published papers on International Conference on Very Large Data Bases, IEEE International Conference on Data Engineering, and IEEE International Conference on BioInformatics and BioEngineering.



Senzhang Wang was born in Yantai, China, in 1986. He received the M.Sc. degree in Southeast University, Nanjing, China, in 2009. He is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, Beihang University, Beijing, China.

His current research interests include data mining and social network analysis.

Zhoujun Li (M'07) received the M.Sc. and Ph.D. degrees in computer science from the National University of Defence Technology, Changsha, China, in 1984 and 1999, respectively.

He is currently with the School of Computer Science and Engineering, Beihang University, Beijing, China, where he has been the Professor, since 2001. His current research interests include data mining, information retrieval, and database. He has published over 150 papers in international journals and conferences, such as the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, *Information Science*, *Information Processing and Management*, World Wide Web Journal, ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2014, SIAM International Conference on Data Mining 2014, ACM International Conference on Information and Knowledge Management 2014, AAAI 2013, and SIGIR 2013.

Xueqiang Lv received the Ph.D. degree in computer science from Northeastern University, Shenyang, China.

He is a Professor with Beijing Information Science and Technology University, Beijing, China. His current research interests include Chinese information processing and multimedia retrieval.