Perspective Hierarchical Dirichlet Process for User-Tagged Image Modeling

Xin Chen1  Xiaohua Hu1  Yuan An1  Zunyan Xiong1  Tingting He2  E.K. Park3
1College of Information Science and Technology, Drexel University, Philadelphia, PA 19104, USA
2Dept. of Computer Science, Central China Normal University, Wuhan, China
3California State University - Chico, Chico, CA 95929, USA

bruce.chen@drexel.edu  thu@ischool.drexel.edu  yuan.an@ischool.drexel.edu
zx26@drexel.edu  tthe@mail.ccnu.edu.cn  ek.uk.park@gmail.com

ABSTRACT
In this paper, we proposed a perspective Hierarchical Dirichlet Process (pHDP) model to deal with user-tagged image modeling. The contribution is twofold. Firstly, we associate image features with image tags. Secondly, we incorporate the user’s perspectives into the image tag generation process and introduce new latent variables to determine if an image tag is generated from user’s perspectives or from the image content. Therefore, the model is capable of extracting both embedded semantic components and user’s perspectives from user-tagged images. Based on the proposed pHDP model, we achieve automatic image tagging with users’ perspective. Experimental results show that the pHDP model achieves better image tagging performance compared to state-of-the-art topic models.

Categories and Subject Descriptors
I.2.6 [Artificial Intelligence]: Learning – Parameter learning; H.2.8 [Database Management]: Database applications – Data mining; Image databases; H.1.2 [Models and Principles]: User/Machine Systems – Human factors, Human information processing

General Terms
Algorithms, Experimentation, Human Factors, Design.

Keywords
Image tagging, Probabilistic generative model, Hierarchical Dirichlet process, User perspective modeling.

1. INTRODUCTION
The prevalence of digital imaging devices, such as digital cameras and digital video cameras, has brought an increasingly large amount of unlabeled multimedia data, especially unlabelled image data. To face the challenge of enormous explosion of unlabeled online image resources, it is important to achieve automatic image tagging for online image resource users. The desirable image tagging system should not only be able to interpret the image content but also be able to integrate users’ contextual information. Breakthroughs in automatic image tagging algorithms will help with organizing the massive amount of online image resources, promote developing and studying of image storage and retrieval systems, and serve for applications such as interest sharing among online image resource users.

Due to its social annotation nature, Flickr image tags have various functional purposes [6]. For example, the topic tags may refer to any object or person displayed in the picture, such as sky, lake, plant life; the time tags indicate the time when a picture was taken; the location tags provide information about sights, like which country it is from; the type tags include camera settings and photographic styles; the usage context tags suggest the context the picture was collected in, while the self reference tags contain highly personal information for the tagger himself, such as ‘diamond class photographer’. Study on different tag categories suggests that topic and location are two most intensively used tag categories in Flickr.com [6]. Further study in social tagging behavior suggests that the factual tags [7] (or the first five tag categories identified in [6]) are more closely related to resource content, while the subjective tags and personal tags are more influenced by users’ perspectives. Generally speaking, compared to subjective and personal tags, factual tags are more relevant to image content. The subjective and personal tags, on the other hand, are usually free-form texts, but they also provide valuable contextual information of users’ tagging preference which can be utilized to customize automatic image tagging for different users.

It should be noted that, some factual tags (which carry on the semantic meanings) are not visible from image contents. For example, the image content usually provides little clues about its location and it is not easy to infer the camera settings and photographic styles. Therefore, in automatic image tagging, one challenging problem is how to bridge over the “semantic gap” [4] between image features and high-level semantic meanings. Specifically, it requires identifying a set of image features that well preserve the semantic consistency of image content. It also requires using generative probabilistic models to infer the less visible semantic concepts (like locations) from the visible parts.

The Correspondence Latent Dirichlet Allocation (CorrLDA) model [1] has been intensively used to model image features with multiple types of associated semantic entities (such as tags) [8, 9]. However, the CorrLDA model requires specifying the exact number of mixture components. In real-world applications, the number of semantic components in an image is unknown. For example, a picture of clear blue sky tend to have less semantic components than an image showing a crowd of people in the street. Therefore, the Hierarchical Dirichlet Process (HDP) model [5], a nonparametric extension of the LDA-based topic models, was proposed. It enables us to represent image content with unbounded number of semantic components, thus provides the...
flexibility of model images with dramatically different contents. In this paper, we extend the HDP model to deal with both users' perspective and the semantic components derived from image contents. We name the new model as 'perspective Hierarchical Dirichlet Process (pHDP) model' as it integrates the user's perspectives into the image tag generation process. The pHDP model introduces new latent variables to determine if an image tag is generated from user's perspectives or from the image contents. Experimental results show that the pHDP model not only generates useful information about semantic components and user perspectives from tagged images, but also achieves better performance in the task of automatic image tagging compared to state-of-the-art topic models.

The remainder of this paper is organized as follows. In Section 2, we review related works in generative topic models. In Section 3, we present the generative process of the proposed pHDp model. Section 4 provides the collapse Gibbs sampling algorithms for model estimation. Section 5 reports the experimental results of the automatic image tagging and compares our approach with several existing models. We conclude the paper in Section 6.

2. BACKGROUND AND RELATED WORK

2.1 Generative Models for Image Tagging

On automatic image tagging, one major task is to identify semantic mixture components from the co-existing image content and text descriptions. In the data mining and information retrieval community, there has been a long time focus on using probabilistic topic models to study the correlation between image and text descriptions. Specifically, the Correspondence LDA (CorrLDA) model [1], which imposes correspondence between text word and other semantic entities, provides a natural way to learn latent semantic components (topics) from image features and associate them with text descriptions. Many recent studies, including sophisticated topic models that associate image features with multiple types of semantic entities (such as protein entities [8], ontology-based biomedical concepts [9]), still follow a similar generative process to the prototype CorrLDA model. In CorrLDA model, each image document has different distribution over semantic mixture components; this feature provides the model a flexibility of adapting to different image contents. However, the CorrLDA model requires specifying the exact number of mixture components, which is fixed for each image document and remains unchanged during the model estimation. In practice, in order to get an optimal number, the researchers have to try out different mixture components numbers and make a choice by comparing the log-likelihood, perplexity and other criteria that indicate how good the model fits the data. The Hierarchical Dirichlet Process (HDP) model [5], is a nonparametric extension of the Latent Dirichlet Allocation (LDA)-based topic models, it enables modeling documents with countable infinite mixture components, thus provides the flexibility of modeling images whose actual semantic component numbers are unknown.

2.2 Modeling User’s Perspective

Study of social tagging in web-based applications has gained increased popularity in the data mining community. Specifically, several probabilistic generative models have been proposed to study users’ tagging patterns [10, 11]. In [11], a topic-perspective (TP) model is proposed to infer how both users’ perspective and the resource content relate to the generation of social annotations. It improves the generative process of social annotations by separating the tag generation process from the generation process of the resource content. While the resource content (such as text words) is only generated from resource topics, the social tags are generated by both resource topic and user perspective. In this model, the user perspective not only refers to the user’s interest, but also covers the user’s expertise, motivation, language and other personal factors.

3. MODELING USER-TAGGED IMAGES

In this section, we introduce the perspective HDP (pHDP) model for user-tagged images. We present graphical representation of pHDP model in Fig. 1. Following the convention in depicting graphical representation of topic models, we use round nodes to represent random variables, in which the white nodes stand for latent random variables, while the gray nodes denote observed ones during the model training. The rounded boxes are used to represent fixed hyper-parameters of the model, while the edges illustrate the conditional dependency in the generative process.

For clarity, we name each tagged image as a document. Some notations to be used in the two models are listed as follows: \( j \) is the number of image documents, \( K \) and \( K' \) (both are countable infinite) indicate the number of semantic mixture components; when \( K \) is a finite number, the models become LDA-like models. To represent the image content, we utilize the saliency features (including visual code-words [12] and MSER feature [13]) as a complement part of the holistic GIST features [3]. Our motivation comes from the fact that the mechanism of human visual perception allows for very rapid holistic image analysis to provide a coarse context of image scene (special layout model), yet it also gives rise to a small set of candidate salient locations in a scene (saliency model) that needs to be intensively studied [2]. In Fig. 1, \( N_j^t \) is the number of tags in document \( j \), while \( N_j^r \) and \( N_j^t \) represent the total number of extracted visual code-words and MSER regions in document \( j \), respectively. In the model, the holistic representation of an image is replicated 10 times to enable the posterior sampling, so \( N_j^h \) denoted the \( h \)th replication of the holistic image representation in document \( j \). In both models, we assume fixed value for Dirichlet process concentration parameters \( a_0 \) and \( \gamma \). We also assume symmetric priors \( a_0, \xi, \zeta, \eta, \xi \) and \( \zeta \) for Dirichlet distributions in the models. Detailed explanations of notations in following discussions are summarized in Table 1.
Table 1. Notations in proposed topic model

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(t, p, v, r, h)</td>
<td>Instances of variables: (t) for tags, (p) for user’s perspective, (v) for visual word, (r) for MSER feature, (h) for GIST feature</td>
</tr>
<tr>
<td>(J, T, U, L)</td>
<td>Number of documents, tags, users, user’s perspectives</td>
</tr>
<tr>
<td>(\omega, g, \delta, s_j)</td>
<td>Indicators for semantic components.</td>
</tr>
<tr>
<td>(K, K')</td>
<td>The number of components at a certain time point.</td>
</tr>
<tr>
<td>(N_{tj}^V, N_{tj}^N)</td>
<td>Number of visual words, MSER regions and tags, plus the replication number of GIST features in document (j).</td>
</tr>
<tr>
<td>(C_{kH}^{★V}, C_{kH}^{★N})</td>
<td>Number of times visual word (v) was assigned to semantic component (k), with/without counting the current instance</td>
</tr>
<tr>
<td>(C_{kH}^{\text{SU}}, C_{kH}^{\text{HS}})</td>
<td>Number of MSER feature vectors being assigned to component (k), with/without counting current instance.</td>
</tr>
<tr>
<td>(n_{k'}^{\text{PO}})</td>
<td>The number of times that perspective (p) is adopted by user (a), except current instance;</td>
</tr>
<tr>
<td>(\bar{n}_{k'}^{\text{PO}})</td>
<td>The number of tags in document (d=j) generated from user’s perspectives ((a=2)) except current instance.</td>
</tr>
<tr>
<td>(n_{k-j}^{\text{PO}})</td>
<td>The number of tags in document (d=j) generated from semantic components ((\gamma=0.1)), except current instance.</td>
</tr>
<tr>
<td>(\alpha_{k'}, \gamma)</td>
<td>Concentration parameters of Dirichlet process</td>
</tr>
<tr>
<td>(\phi_1^k, \phi_1^{k'})</td>
<td>The tag distribution of semantic component (k, k').</td>
</tr>
<tr>
<td>(\phi_\gamma)</td>
<td>The visual word distribution of semantic component (k)</td>
</tr>
<tr>
<td>(\pi_k)</td>
<td>The document-level weights of semantic component indicators for document (j).</td>
</tr>
<tr>
<td>(\alpha_{k'}, \beta, \gamma_{\text{tr}}, \eta_{\text{tr}}, \xi_{\text{tr}})</td>
<td>Hyper-parameters of Dirichlet distributions</td>
</tr>
<tr>
<td>(\mu^{\text{G}}, \sigma^{\text{G}})</td>
<td>Parameters of the (n)th Gaussian distribution with respect to the (k)th semantic component</td>
</tr>
<tr>
<td>(\overline{\pi}<em>{k,j}, s</em>{k,j}^2)</td>
<td>Sample mean and sample variance of the (n)th Gaussian distribution with respect to the (k)th semantic component</td>
</tr>
<tr>
<td>(\theta_{\text{tr}})</td>
<td>The perspective distribution of user (a), and the tag distribution of perspective (p).</td>
</tr>
<tr>
<td>(x, \lambda_j)</td>
<td>Switch variable that decides the source of each tag and the document-level distribution of different (x) values</td>
</tr>
<tr>
<td>(\beta)</td>
<td>The global weight of semantic component indicators across the corpora</td>
</tr>
</tbody>
</table>

As shown in Fig. 1, this model primarily comprises of two parts split by the dash line. The part on the right hand side is essentially the standard HDP model. The generative process of this part begins with drawing a global probability measure \(G_0 \sim \text{DP}(\pi, H)\) and for each document \(j\), draw a child Dirichlet process \(G_j \sim \text{DP}(\alpha_0, G_0)\). Following the stick-breaking construction, it is equivalent to firstly drawing a global weight \(\beta \sim \text{GEM}(\gamma)\) for semantic component indicators \(k\), then for each document \(j\), draw the document-level weights of semantic component indicators \(\pi_j \sim \text{DP}(\alpha_0, \beta)\). The data observations in document \(j\) are generated by repeatedly drawing semantic component indicator \(z_j\) and \(z_j\) from \(\pi_j\) and then draw each data observation (i.e. each MSER region and each visual code-word) from the conditional probability of the sampled semantic component. The left half of the model is for the generation of image tags. As mentioned in Section 1, image tags have various functional purposes. For example, some tags (like most factual tags) are closely related to the contents displayed in images, while other tags (including location tags, subjective tags and personal tags) indicate user’s contextual information as well as his/her subjective feelings and preferences, which we refer to as “user’s perspectives”. Accordingly, the generative process of user-tagged images should be able to take into account both user’s perspectives and semantic components from image contents. In pHDP model, each tag \(t\) created by user \(u\) for document \(j\) can be either drawn from the semantic components associated with \(j’s\) image content or from \(u’s\) perspectives. To decide the source of each tag, a switch variable \(x\) is introduced. For each tag \(t\) in document \(j\), the value of \(x_j\) (which takes values 0-2) is sampled from a multinomial distribution \(x_j\) (with a Dirichlet prior \(\zeta\)). When the value of \(x_j\) equals 0 or 1, the topical indicator of tag \(t\) is drawn uniformly from the semantic components learned from the image contents (the red dashed arrows in Fig. 1 show this process). When \(x_j\) equals 2, a user perspective \(p\) will be sampled from the perspective distribution \(\theta_u\) for user \(u\), and tag \(t\) will be drawn from the tag distribution \(\eta_p\) of perspective \(p\) (the blue arrows in Fig. 1 illustrate this procedure). The switch variable \(x\) plays a critical role in the pHDP model; it is a personalized factor showing to which extent the user perspective influence the tagging result. It provides the model a flexibility to determine if a specific image tag relates to the semantic components displayed in an image, or it relates to user’s context information as well as his/her subjective feeling and preference (user’s perspective). The generative process of the pHDP model is represented in Table 2.

Table 2. The generative process of proposed models

1. Draw a global weight \(\beta \sim \text{GEM}(\gamma)\);
2. For each semantic component \(k\), draw \(\lambda_k \sim \text{Beta}(1, \zeta)\) , \(\phi_\gamma \sim \text{Dirichlet}(\zeta)\), \(\phi_\gamma \sim \text{Dirichlet}(\zeta)\);
3. For each semantic component \(k\), sample Gaussian-parameters \(\mu_{\lambda_k}, \sigma_{\lambda_k}, \mu_{\phi_\gamma}, \sigma_{\phi_\gamma}\) from sample mean and sample variance;
4. For each user \(u\), sample \(\theta_u \sim \text{Dirichlet}(\alpha_u)\), for each user perspective \(p\), sample \(\psi_p \sim \text{Dirichlet}(\eta)\);
5. For the \(j\)th document, draw \(\pi_j \sim \text{DP}(\alpha_0, \beta)\), \(\pi_j' \sim \text{DP}(\alpha_0, \beta')\);
6. For the holistic scene representation of the \(j\)th tagged image ,
   a. sample scene component indicator \(s_j \sim \text{Discrete}(\pi_j)\),
   b. for the \(n^\text{th}\) dimension of the GIST feature vector \(h_j\),
      i. sample \(h_{\text{tr}} \sim N(\mu_{\theta_{\text{tr}}}^{\text{tr}}, \sigma_{\theta_{\text{tr}}}^{\text{tr}})\)
7. For the \(j\) of the \(N_j\) visual code-words in the \(j\)th document
   a. sample object component indicator \(z_j \sim \text{Discrete}(\pi_j)\)
   b. sample its texton id \(v_j \sim \text{Multinomial}(\phi_v)\);
8. For the \(j\) of the \(N_j\) MSER salient regions in the \(j\)th document
   a. sample object component indicator \(z_j \sim \text{Discrete}(\pi_j)\)
   b. for the \(n^\text{th}\) dimension of MSER feature vector \(r_{ji}\),
      i. sample \(r_{ji} \sim N(\mu_{\theta_{\text{tr}}}^{\text{tr}}, \sigma_{\theta_{\text{tr}}}^{\text{tr}})\)
9. For each document \(j\), sample \(\lambda_j \sim \text{Dirichlet}(\zeta)\);
10. For each tag \(t\) in document \(j\) created by user \(u\);
   a. sample a switch variable \(x \sim \text{Multinomial}(\lambda_j)\);
   b. if \((x = 0)\)
      Sample semantic component indicator \(z_t \sim \text{Discrete}(\pi_t)\)
      Sample a tag \(t \sim \text{Multinomial}(\phi_t)\)
   c. if \((x = 1)\)
      Sample semantic component indicator \(z_t \sim \text{Uniform}(\pi_t')\)
      Sample a tag \(t \sim \text{Multinomial}(\phi_t')\)
   d. if \((x = 2)\)
      Sample a perspective \(p \sim \text{Multinomial}(\theta_p)\)
      Sample a tag \(t \sim \text{Multinomial}(\psi_p)\).
4. GIBBS SAMPLING FOR THE MODEL ESTIMATION

In this section, we describe the Gibbs sampling scheme for the proposed pHDP model. The sampling scheme consists of two steps. The first step is sampling for semantic component indicators $z$ as well as the corresponding HDP hyper-parameters $\beta$. In order to sample a HDP-like model, one may either follow the Chinese restaurant franchise (CRF) or use direct assignment [5]. In our work, the direct assignment is used (Table 3).

**Table 3. The posterior sampling of semantic components**

**Preliminaries:**
Suppose that at current stage of the sampling, only $K$ of $L \rightarrow \infty$ semantic components have been assigned to the observations, define: $\beta = \sum_{k=1}^K \beta_k$, and $\gamma = \gamma / L$, $\gamma = \gamma (L-K) / L$, then we get:

$$\beta = \{ \beta_1, \ldots, \beta_k, \beta_u \} \sim \text{Dirichlet}(\gamma, \ldots, \gamma, \gamma).$$

**Repeat for each data observation until convergence:**

**Sampling $z$ (may equates to an existing $k$ or $k_{new}=K+1$):**
Firstly, integrate out $\pi_j$ to get the marginal probability $p(z|\beta)$:

$$p(z_j = k | v_j, z^p, v^p, \beta) \propto p(z_j = k | z^p, \beta) p(v_j | z_j, z^p, v^p, \beta) \propto \left(\begin{array}{c} \alpha_k \beta_k + n_{j,k} \end{array}\right) f_{s,v_j}^{-1}(v_j) \left(\begin{array}{c} \alpha_k \beta_k f_{s,v_j}^{-1}(v_j) \end{array}\right)$$

For visual word $v^p$, $f_{s,v_j}^{-1}(v_j) = \sum_{v \in V} f_{s,v_j}(v) / \sum_{v \in V} f_{s,v_j}(v) = \propto \frac{\xi_v}{\sum_{v \in V} \xi_v}$

Similarly, for both MESR feature vector and GIST feature vector,

$$f_{s,v_j}^{-1}(v_j) = \sum_{v \in V} f_{s,v_j}(v) / \sum_{v \in V} f_{s,v_j}(v) = \propto \frac{\xi_v}{\sum_{v \in V} \xi_v}$$

in which $t_{n+1}(u, s^2 / n)$ denotes the student-t density with mean $u$, variance $s^2 / n$ and $n-1$ degree of freedom.

**Sampling $m$ (feasible when $n_{j,k} < 200$):**
For each $j$, the auxiliary variable $m (0 \leq m \leq n_{j,k})$ is sampled as:

$$p(m_j = m | m^k, z, \beta) = \frac{\Gamma(\alpha_k \beta_k)}{\Gamma(n_{j,k} + \alpha_k \beta_k)} \binom{n_{j,k}}{m} (\alpha_k \beta_k)^m$$

in which $s(n,m)$ is defined as: $s(0,0)=s(1,1)=1, s(n,0)=0, s(m,0)=0, m>n$, $(n+1,m)=s(n-1,m)+ns(n,m)$

**Sampling $\beta$:** accumulate $m_k$ for all document $j$ to get $m_j, m_{z}, \ldots, m_{K}$, then draw $\beta \sim \text{Dirichlet}(m_j, m_z, \ldots, m_K, \gamma)$.

The second step is sampling for switch variable $x$, perspective distribution $\theta_{x}$, tag distribution $\phi_k$, $\phi_k^{*}$ and $\psi$. We derive the sampling equation of switch variable $x_j$ for each tag $t=q$ in document $d=e$ as follows:

$$p(x_j = 0, z_j = k | t = q, z_{-j}, t_{-q}, \xi_v) \propto \frac{n_{j,q} + \xi_v}{n_{j,q} + n_{j,q} + 3 \xi_v} C_{ik}^{iz} + C_{ik}^{iz} + C_{ik}^{iz}$$

$$p(x_j = 1, z_j = k | t = q, z_{-j}, t_{-q}, \xi_v) \propto \frac{n_{j,q} + \xi_v}{n_{j,q} + n_{j,q} + 3 \xi_v} C_{ik}^{iz} + C_{ik}^{iz} + C_{ik}^{iz}$$

$$p(x_j = 2, z_j = p | t = q, z_{-j}, t_{-q}, \xi_v) \propto \frac{n_{j,q} + \xi_v}{n_{j,q} + n_{j,q} + 3 \xi_v} C_{ik}^{iz} + C_{ik}^{iz} + C_{ik}^{iz}$$

5. EXPERIMENTS AND RESULTS

In this section, we investigate the performance of the proposed pHDP model in automatic image tagging experiments using the MIR-Flickr dataset, which is composed of 25000 images covering a wide spectrum of image categories (contributed by a total of 9862 Flickr users). In total, there are 302447 tags, with a vocabulary size (number of unique tags) of 64037; thus the average number of tags per image is 8.94. In the image tagging experiment, we use a 50% subset of the MIR-Flickr collection as training data and the other 50% as testing data (with tags removed). On constructing the two subsets, we ensure that tagged images from the same user are equally split to both subsets. The values of global concentration parameter $\gamma$ and the user perspective parameter $\lambda$ are determined by perplexity comparison on a serial of values. Other hyper-parameters (such as Dirichlet distribution priors: $\alpha_k, \xi_v, \xi_k, \eta$ and $\zeta$) are set in prior and fixed during the experiments. The prediction of image tags for the testing images is achieved by performing another Gibbs sampling on testing images to estimate the document-level distribution of switch variable and semantic components, with a fixed set of semantic components and user perspectives estimated from the training dataset. On the convergence of Gibbs sampling, the probability of tagging an image $j$ from user $u$ with tag $t_j$ is:

$$p(t_j | u) = \frac{1}{K} \sum_{k=1}^K p(t_j | z_k) p_{ttag}(z_k | j) + \frac{1}{K} \sum_{k=1}^K p(t_j | z_k) p_{ttag}(p_j | u)$$

The performance is evaluated by perplexity and tagging accuracy.
5.1 Perplexity Comparison

The perplexity is a standard criterion for generative probabilistic models that evaluates how well the model predicts the testing data. The perplexity of a testing image dataset $D_{test}$ is:

$$\text{perplexity}(D_{test}) = \exp \left[ \frac{-\sum_{j=1}^{N} \log(p(t_j))}{\sum_{j=1}^{N} N_j} \right]$$

The perplexity score for a model is the lower the better.

Fig. 2(a) shows the perplexity changing of the proposed pHDP model and baseline models (CorrLDA model and HDP model) over the iterations during the Gibbs sampling process. We test pHDP model on a serial of $\gamma$ values. For CorrLDA model and HDP model, we only show their perplexity scores under the optimal parameter settings (i.e. CorrLDA model with 75 topics and HDP model with $\gamma = 15.0$). The results show that pHDP model achieve best performance with $\gamma = 15.0$ and it outperforms both HDP model and CorrLDA model. Fig 2(b) represents the perplexity of pHDP model under different perspective numbers. The optimal perspective number is L=75.

5.2 Image Tagging Accuracy Evaluation

Using Eq. (7), we calculate the probability of tagging an image $j$ from user $u$ with different tags. Tags with highest probability are used for tagging. After that, the predicted top-ranked image tags are compared with the ground truth for validation. If a predicted tag finds exact match in the ground truth tags, it will be considered as one hit. The ratios of hit numbers over the predicted tag numbers are averaged to produce the final annotation accuracy.

Fig. 3 illustrates examples of image tagging results. Fig. 3(a) is an image shows a winter night in Toronto, Ontario, Canada. The ground truth image tags involve both location tags (‘ontario’, ‘canada’) and topic tags (such as ‘clouds’, ‘lake’, ‘night’, sky and ‘water’). However, the image content alone provides little clues about the location. Further studies indicate that other images contributed by the same user are also tagged with ‘ontario’ and ‘canada’. This may suggest that the user lives in Ontario, Canada.

During the pHDP modeling, this user contextual information is captured as a part of the user’s perspectives. When tagging a new image from the same user, the pHDP model will smooth the document-level predictive tag distribution with user’s perspective and allow for tagging with location tags (Fig. 3(a) and 3(b), highlighted in bold). Fig 3(c) is contributed by a user from Malaysia. Similarly, the user contextual information is captured in user’s perspectives. Thus the pHDP model succeeds in tagging image with both location tags (such as ‘malaysia’) and type tags (camera settings, like ‘nikon’).

As shown in Fig. 3(c), tags predicted by the pHDP model also involve subjective tag, like ‘interesteness’, which demonstrates that the pHDP model is also capable of modeling user’s subjective feelings. As a comparison, the HDP model fails to predict either location tags or subjective tags since it only relies on image content to make tag predictions.

User ID: 17875539@N00
Title: City with ice
Tags: lake, night, ontario, toronto, torontoharbour, canada, ice, frozen, winter, dawn, morning, nature, sky, water, landscape, snow, nikon, nikond200, cold, landscape, lake, water, outdoor, outdoorphotography, frost, frosty

Top ranked tag Probability
sky 0.0329
night 0.0265
transport 0.0230
clouds 0.0209
water 0.0173
sea 0.0102
road 0.0097

tags predicted by pHDP
(a) tagging results of image entitled ‘City with ice’

User ID: 17875539@N00
Title: Some say in ice
Tags: lake, night, ontario, toronto, torontoharbour, canada, ice, frozen, winter, dawn, morning, nature, sky, water, landscape, snow, nikon, nikond200, cold, cherrybeach, outdoor, natural, frost, frosty

Top ranked tag Probability
sky 0.0375
structures 0.0306
water 0.0199
sea 0.0173
flower 0.0162
transport 0.0139
outdoor 0.0124

tags predicted by pHDP
(b) tagging results of image entitled ‘Some say in ice’
8. REFERENCES


6. CONCLUSIONS

In this paper, we proposed a perspective Hierarchical Dirichlet Process (pHDP) model to deal with user-tagged image modeling. The contribution is twofold. Firstly, we associate image features with image tags. Secondly, we incorporate the user’s perspectives into the image tag generation process and introduce new latent variables to determine if an image tag is generated from user’s perspectives or from the image content. Therefore, the model is capable of extracting both embedded semantic components and user’s perspectives from user-tagged images. Based on the proposed pHDP model, we achieve automatic image tagging with users’ perspective. Experimental results show that the pHDP model achieves better image tagging performance compared to state-of-the-art topic models.

7. ACKNOWLEDGMENTS

This work was supported in part NSF CCF 0905291, NSF CCF 1049864, NSFC 90920005.