2013 IEEE International Conference on Systems, Man, and Cybernetics

Discovering User Preference from Folksonomy

Xiaohui Guo, Richong Zhang, Jinpeng Huai, Hailong Sun, Xudong Liu
Institute of Advanced Computing Technology, School of Computer Science and Engineering
Beihang University
Beijing, China
Mail:{guoxh, zhangrc, sunhl, liuxd}@act.buaa.edu.cn

***Abstract*—The increasing availability of socially shared media with tags annotated makes it vital for retrieval approaches to precisely detect web content topic semantic and better understand user interest. Most existing methodologies process the queries merely considering user posted keywords and retrieve media labeled with tags that are similar to query words, while ignoring users implicit interests and preferences. This fact stimulates us to develop preference discovering models to reveal the users' latent intents. In this paper, we study the problem of finding user preference and interest from folksonomy corpus and propose a preference-topic model that exploits probabilistic graphical model and Gibbs sampling algorithm to infer the user interested latent semantic topics. The experimental results show that, with the help of the proposed model, preference topics of the web content creators can be effectively discovered. In addition, two exemplified applications are discussed briefly.**

***Index Terms*—Folksonomy, Social Tagging, Interest, Preference Discovery**

I. Introduction

Acquiring and recognizing the preference of social media users is an appealing functionality of the online social communities and intelligent services. Such information is typically not directly available, and often underlies in the media that the users have created and gleaned bit by bit. Algorithmically analyzing the content of web digital media, such as photos, videos etc., directly from the visual or low level features to capture the high level semantic information like the users' interest is a complex and difficult task. Meanwhile, the massive media volume in current big data era is doomed to require sophisticated and costly computational infrastructure. However, many vertical online social communities urgently need to target the group of users with similar interests to promote the unpopular contents or media hibernating in their repositories. On the other hand, the individual users are usually at a loss to face so many unexploited contents mismatching their intent, and need some effective and smart systems to return the desired and personalized information with less effort. For example, before the journey, while exploring the intended destination on a travel community, the traveler wish the system to intelligently render some scenic spots, seasonal events, and local snacks which are cater to his preferences and interests.

Folksonomy data (also known as social tagging or bookmarking), formed by the crowds to category and index their favoured web contents through annotating tags on them, pave an alternative and convenient way to cope with the above challenges. The individual tag seems with less value, but all the user generated tags of the whole community could collaboratively and statistically emerge significant semantic structure. This harvested collaborative intelligent information could be utilized as a kind of rich, diverse, and sufficient description of the media itself. For instance, only from the tags associated on a Flickr photo, we could achieve many information, such as the people or things involved, location shoot at, content category, even quality, and devices used. Furthermore, labeling tags on a web content is a convincing indicator of the user interest and preference to the media content. In another word, the personomy, a personal view of some folksonmy, is a informative information source to mine the user preference, e.g., user GraemeNicol's tagcloud[1] on flickr. All of these undoubtedly provide an opportunity to the social media communities with the folksonomy facility intelligently comprehending the users' preferences and interests.

To address this problem, we propose a probabilistic graphical model, called *Preference-Topic model*, for the discovery of the preference of the photo creators in Flickr sites, by modeling the preference as a latent variable to bridge the gap between the photo creators and the tags associated with photos. Specifically, the preference of a photo creator is modeled via a multinomial distribution over the set of topics and the set of tags annotated on a photo are modeled as being drawn independently from the a multinomial distribution depending only on the latent topics. These two multinomial distributions are assumed as being drawn from two Dirichlet distributions prior respectively. We derive a Gibbs Sampling algorithm for the inference of these two multinomial posterior distributions. The derived perceptible topical semantic structure (e.g. animal, sport, food, landmark, landscape, etc.) experimentally verify that our model and inference algorithm is capable of effectively identifying the user preferences as well as the probability of tags associated with any given topic. In addition, we employ the cosine similarity as a metric to measure the similarity between users' preference distributions, and give a matrix visualization of user clustering result to demonstrate the usage of such model.

This paper is organized as follows. The next section describes related works. Section 3 first introduces the Preference-Topic model and gives the corresponding Gibbs sampling algorithm. Section 4 presents the experimentally discovered hidden preference topics, and two practical scenarios are introduced to illustrate the application potential of our model.

[1] http://www.flickr.com/photos/slavers/tags/

978-1-4799-0652-9/13 $31.00 © 2013 IEEE
DOI 10.1109/SMC.2013.362

II. Related Work

Incorporating social tagging, media searching and recommending has been enhanced greatly in this web 2.0 age. Relevant research works mainly consist of tagging based query reformulation and expansion [12], improving the semantic annotation quality of folksonomy corpus by refinement[14], re-tagging and tag recommendation [7] etc. Tags was utilized as an important textual feature by social media systems. These studies mainly focus on finding proper tags for social media contents, however, only a few researches consider the user's preference and interest discovery from social tagging systems.

Probabilistic topic models has been widely used to express the hidden thematic structures for specific applications about text documents [8], visual images and videos [12], even human activities recognition [2] and social network analysis [9]. Except the LSA extracting topic by SVD decomposition, pLSA [5] and LDA [1] both model each item of a collection as a mixture of topics (i.e., probability distribution over topics), and each topic as a probability distribution of words. Author-Topic model [11] inherits the topic's representation, and assume the author as a mixture of topics. Through Place Dirichlet prior on the model parameters, LDA and Author-Topic model usually adopt bayesian inferene to estimate the posterior distribution. However, pLSA usually resorts to EM algorithm to estimate the parameters. Inspired by the bag of words assumption and the conjugacy between multinomial and Diricihlet distribution, we custome our model specific to folksonomy dataset and characterize the dependencies ammong users, photos, tags and topics.

Preference learning [3] utilizes ordering relation to model the possible choices, but we advocate the probabilistic latent factor model to infer the preference distribution on the low dimensional semantic space derived from the folksonomy.

III. Preference Discovering Model

In this study, we take tags as sufficient and yet noisy descriptions of the annotated photos, and assume the photo annotated with tag as a mixture of photo creator's preference topic. The Preference-Topic model, a probabilistic generative model, for the users' preference discovery problem in folksonomy corpus is fomulated in the following subsection. A Bayesian inference method, Gibbs sampling algorithm, is also derived.

A. Preference-Topic Model

We use D to denote the set of all observed photos and d_i to denote the i^{th} photo in D. Likewise, we use A to denote the set of all photo creators and $a_d \in A$ denotes the creator for the photo $d \in D$. We denote by W the set of all observed tags (aka, the tokens or placeholders for tags), and by W_d the set of tags annotated on photo $d \in D$. And $\vec{\mathbf{w}}$ denote the vector variable of tag tokens in the corpus.

Similar to the traditional topic models, we introduce a latent variable z to assign topic to tags, which takes on values in a finite set $T := \{1, 2, \ldots, K\}$, where K is the topic dimension hyperparameter. $\vec{\mathbf{z}}$ is the topic assignment vector corresponding to and with same dimension to tag tokens vector $\vec{\mathbf{w}}$. And $z_d^i = k, k \in T$, means assigning topic k for the i^{th} tag $w_d^i \in W_d$. In another word, w_d^i will be drawn from the tags distribution in the k^{th} topic in our latter discussed generative process.

For each photo creator $a \in A$, we model their interest and preference probabilistically, where we use $\theta_{a,t}$ to denote the probability, $p(t|a)$, that user a favours topic $t \in T$. Collectively, We treat $\{\theta_{a,t} : a \in A, t \in T\}$ as a random matrix denoted by Θ. In addition, we assume that for each $a \in A$, θ_a ($\Theta_{a\cdot}$, the a^{th} row of Θ) is drawn from Dirichlet prior distribution $Dir(\alpha)$, which is the preference topic multinomial distribution over the K latent topics. For the symmetric hyperparameter α, empirically, we should set it less than 1, so that the majority of the probability mess shrink to the corner of the $K-1$ simplex. Thus, our model should achieve better discrepancy of users' preference topics.

Furthermore, we assume that a specific tag $w \in W$ has multinomial probability $p(w|t)$ emerging in some topic $t \in T$, and intuitively, the higher probability indicates w is more expressive tag in the topic t. Similarly, we use matrix $\Phi = \{\phi_{t,w} : t \in T, w \in W\}$ to denote the distributions over tags associated with each topic. And each row ϕ_t ($\Phi_{t\cdot}$, the t^{th} row of Φ) is exactly the probability $p(w|t)$ for each $w \in W$ given a unique topic t, which is also generated from a Dirichlet prior distribution symmetrically parameterized by β, $Dir(\beta)$.

Then the proposed Preference-Topic model is a generative model specifying the tag generation process as follows:

1) For each photo creator $a \in A$, generate $\theta_a \sim Dir(\alpha)$; For each topic $t \in T$, choose $\phi_t \sim Dir(\beta)$.
2) For each photo $d \in D$
   a) based on its creator a_d, draw a preferred topic $z_{a_d} \sim multinomial(\theta_{a_d})$;
   b) based on z_{a_d}, draw the set of tags w_d i.i.d from $multinomial(\phi_{z_{a_d}})$.

The graphical (Bayesian network) representation of this model is shown in Figure 1. Treating θ and ϕ as random variables, the model then precisely specifies the joint distribution parameterized by α and β as follows:

$$\begin{aligned} & p(\vec{\mathbf{w}}, \vec{\mathbf{z}}, \Theta, \Phi | A, \alpha, \beta) \\ = \; & p(\Phi|\beta)p(\Theta|\alpha) \prod_{d \in D} \prod_{i=1}^{|W_d|} p(z_d^i|\theta_{a_d}) p(w_d^i|\phi_{z_d^i}) \end{aligned} \quad (1)$$

B. Inference

In many potential application for predicting task, we pursuit the posterior distribution, or the expectation of some function of posterior. For continuous random variables in topic models, like θ and ϕ in our model, there exist a variety of estimation algorithms, including variational inference[1], expectation propagation[10], MCMC[4], and belief propagation[15].

As to our Preference-Topic model, the posterior distribution of model parameters $\{\Theta, \Phi\}$ takes the form

$$\begin{aligned} & p(\Theta, \Phi | \vec{\mathbf{w}}, A, \alpha, \beta) \\ = \; & \int_{\vec{\mathbf{z}}} p(\Theta, \Phi | \vec{\mathbf{z}}, \vec{\mathbf{w}}, A, \alpha, \beta) p(\vec{\mathbf{z}} | \vec{\mathbf{w}}, A, \alpha, \beta) \, d\vec{\mathbf{z}}. \end{aligned}$$

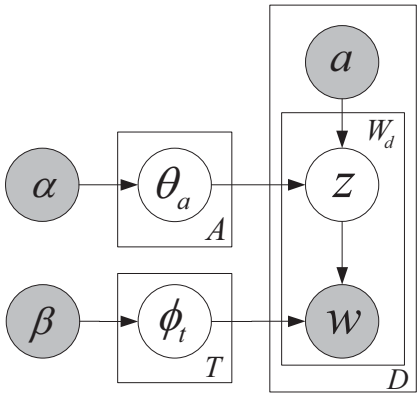

Fig. 1. Graphical representation of Preference-Topic Model

Gibbs Sampling, a form of MCMC inference, is exploited to approximate this posterior Distributions. Firstly, we use Gibbs sampling to achieve a point estimation of $p(\vec{\mathbf{z}}|\vec{\mathbf{w}}, A, \alpha, \beta)$, and then, In the light of the fact that Dirichlet distribution is the conjugate prior for the multinomial distribution, the expectation of Θ and Φ can be straightforward derived.

The joint distribution of generating all observed tags and its topic assignments $p(\vec{\mathbf{z}}, \vec{\mathbf{w}}|A, \alpha, \beta)$ can be obtained by integrating out θ and ϕ from Eq.1. To construct the Markov chain for sampling that converges to the stationary posterior distribution $p(\vec{\mathbf{z}}|\vec{\mathbf{w}}, A, \alpha, \beta)$, we give the Markov transition probability as follows:

$$\begin{aligned} & p(z_d^i = t|w_d^i = w, \vec{\mathbf{z}}^{-di}, \vec{\mathbf{w}}^{-di}; a_d, \alpha, \beta) \\ \propto \; & \frac{{C^{AT}}_{at}^{-di} + \alpha}{\sum_{t'}^{K} ({C^{AT}}_{ut'}^{-di} + \alpha)} \frac{{C^{TW}}_{wt}^{-di} + \beta}{\sum_{w' \in W} ({C^{TW}}_{w't}^{-di} + \beta)} \end{aligned} \quad (2)$$

During the sampling process in our algorithm, we maintain two matrix, creator-topic matrix C^{AT} of size $A \times T$ and topic-tag matrix C^{TW} of size $T \times W$, whose entries stores the corresponding frequency of tags assigned topic t for photo creator a's photo (denoted by $n_a^{(t)}$) and tag w assigned to topic t (denoted by $n_t^{(w)}$), up to the current iteration. ${C^{AT}}_{at}^{-di}$ in above Eq.2 is the number of tag tokens assigned to topic t associated to the a's photos excluding the current assignment of tag w_d^i. Similarly, ${C^{TW}}_{wt}^{-di}$ is also excluding the assignment for w_d^i. Meanwhile $\vec{\mathbf{z}}^{-di}$ and $\vec{\mathbf{w}}^{-di}$ stand for the current topic assignments and tag observations vector respectively, except for the current tag w_d^i.

After the burn-in time, utilizing the conjugacy property and the current sample, we can derive the posterior as following:

$$\begin{aligned} \theta_a|\vec{\mathbf{z}}, \vec{\mathbf{w}}, a, \alpha &\sim Dir(\vec{n_a} + \alpha) \\ \phi_t|\vec{\mathbf{z}}, \vec{\mathbf{w}}, \beta &\sim Dir(\vec{n_t} + \beta) \end{aligned}$$

Where $\vec{n_a}$ is the a^{th} row of C^{AT}, i.e. ${C^{AT}}_{a\cdot}$; $\vec{n_t}$ is the t^{th} row of C^{TW}, i.e. ${C^{TW}}_{t\cdot}$. The Kullback-Leibler divergence could be used to measure the discrepancy between two samples by different iterations, so that we could determine the convergence of Markov Chain Monte Carlo process. Through computing the expectation of posterior distribution, that yields:

$$\theta_{a,t} = \frac{n_a^{(t)} + \alpha}{\sum_{t' \in T}(n_a^{(t')} + \alpha)} \quad \phi_{t,w} = \frac{n_t^{(w)} + \beta}{\sum_{w' \in W}(n_t^{(w')} + \beta)} \quad (3)$$

IV. Experimental Study

In order to demonstrate our model's performances of discovering the preference semantic structure underlying a real-world Flickr data, we downloaded 57853 photos tagged with "Beijing" through Flickr public APIs. There are 1609 users and 18859 tags corresponding to these photos, and we abandon the non-English tags at the data preprocessing stage. The total number of tag occurrences on these photos is 430856.

Specifically, in this section, we firstly give a method to tune the proposed model parameters. Then a detail explanation of our experimental results is presented. Finally, we reveal several potential application scenarios, to manifest the practical significance and usage of model.

A. Tuning the Parameters

Similar to those suggested in [4][11], we have heuristically chosen $\alpha = 50/|T|$ and $\beta = 0.01$ respectively in our Gibbs Sampling algorithm. For the latent topic number hyperparameter, we exploit the *perplexity* to determine it, which is a conventional metric in language model, measuring how well the model fits the data, and lower perplexity correlates with better recognition performance [6]. Formally,

$$Perplexity = \exp\left\{-\frac{\sum_{d \in D} \log p(W_d)}{\sum_{d \in D} |W_d|}\right\} \quad (4)$$

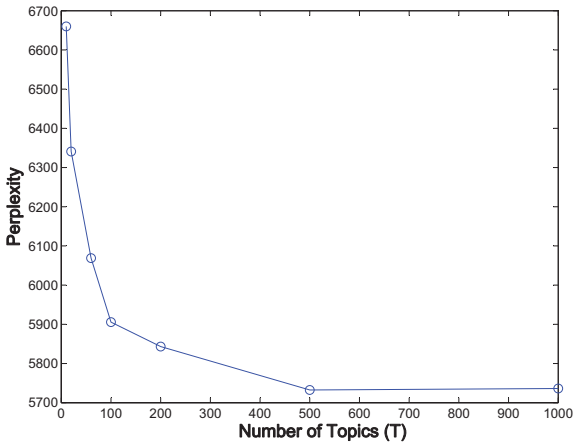

Fig. 2. Perplexity of the experiment corpus, with topic number set to 10, 20, 60, 100, 500, and 1000.

We conduct a series of experiments for different topic numbers. The perplexity values of the entire dataset is computed based on Eq.4 for the given topic numbers and are shown in Fig.2. The perplexity initially decreases as a function of T, reaches a minima at $T = 500$, and then rises slightly thereafter. This suggests that the data is best interpreted by a model incorporating 500 topics. Since the perplexity difference for

topic number between 100 and 500 is small, we chose the topic number 100 to conveniently present the experiment results.

B. Discovering the Latent Preference Semantic Structure

Although the photos collected are only through the tag "Beijing", diverse topics are derived from the tags. After 1500 iterations of our Gibbs sampler, we calculate the prediction posterior probability through Eq. 3. Table I illustrates the discovered most representative 8 latent topics out of totally 100. For each topic, we outline the top 12 most informative tags on the upper rows, and the top 12 most likely photo creators favoring the topic on the under rows respectively.

TABLE I
ILLUSTRATION OF 8 PICKED REPRESENTATIVE PREFERENCE TOPICS

Topic 10		Topic 27		Topic 48		Topic 70	
TAG	**PROB.**	**TAG**	**PROB.**	**TAG**	**PROB.**	**TAG**	**PROB.**
China	0.3146	China	0.1567	food	0.3115	night	0.0815
Pekin	0.0694	macro	0.1342	mannequin	0.0510	Water	0.0373
Shanghai	0.0548	insect	0.1266	seafood	0.0283	sky	0.0362
Shandong	0.0423	Lepidoptera	0.0924	airplane	0.0261	light	0.0342
dvyang	0.0350	moth	0.0545	China	0.0227	lake	0.0319
henan	0.0337	caterpillar	0.0433	yokohama	0.0210	red	0.0298
Everyday	0.0297	larva	0.0420	lunch	0.0171	sunset	0.0278
Gongyi	0.0257	butterfly	0.0321	fish	0.0167	blue	0.0264
Zhengzhou	0.0244	dragonfly	0.0176	sashimi	0.0158	green	0.0262
Weihai	0.0238	green	0.0161	Ginza	0.0120	tree	0.0177
Jinan	0.0218	brown	0.0160	Crab	0.0107	Clouds	0.0173
indoor	0.0170	Limacodidae	0.0129	octopus	0.0090	lights	0.0170
USER	**PROB.**	**USER**	**PROB.**	**USER**	**PROB.**	**USER**	**PROB.**
M. Jean	0.3758	itchydog	0.9233	*DigiPub*	0.7826	m.tomic	0.1298
DvYang	0.0880	R. Lee	0.0012	Ray Yu	0.089	S. Zimny	0.0566
B. Glover	0.0572	C. Harry	0.0010	keeskee	0.0246	Choollus	0.0526
ironde	0.0504	ChicoGoya	0.0010	Dollynpc	0.0223	luzhouzjy	0.0412
amCharlene	0.0367	C. A. Minic	0.0006	ROSS HK	0.0204	PhotonMix	0.0369
GraemeNicol	0.0190	neonbubble	0.0006	beatbull	0.0176	China Chas	0.0330
ada	0.0116	W.S.Boxing	0.0005	aye kaye	0.0106	ROSS HK	0.0321
photogarret	0.0099	jailman	0.0005	avlxyz	0.0099	*ken*	0.02443
C. Lopez	0.0088	J.-J. Pous	0.0005	Choollus	0.0099	Subu-yan	0.0226
A. T. Ochoa	0.00427	Felstone	0.0005	Fiorano	0.0095	B. Lepley	0.0216
Egyptian Mau	0.00313	*DigiPub*	0.0003	AsiaCz	0.0079	J. K. Read	0.0178
le niners	0.0031	Ankhmork	0.0003	marin.tomic	0.0075	ochurchill	0.0155

Topic 80		Topic 84		Topic 86		Topic 95	
TAG	**PROB.**	**TAG**	**PROB.**	**TAG**	**PROB.**	**TAG**	**PROB.**
Beijing	0.9278	palace	0.0917	Beijing	0.2051	film	0.1320
park	0.0118	Summer	0.0837	Honeymoon	0.1409	kodak	0.1150
jingshan	0.0089	wall	0.0786	Olympics	0.0774	Taiwan	0.0664
Beihai	0.0054	forbidden	0.0718	statue	0.0286	rolleiflex	0.0656
Huamao	0.0042	great	0.0666	Sport	0.0271	Taipei	0.0478
qianmen	0.0024	City	0.0558	Stadium	0.0240	minolta	0.0472
Zhengyangmen	0.0018	Heaven	0.0546	cycling	0.0216	autocord	0.0444
Holga	0.0016	temple	0.0482	hutong	0.0193	rollei	0.0351
promenade	0.0015	Mutyanju	0.0251	lake	0.0147	japan	0.0166
video	0.00133	tower	0.0191	Olympic	0.0147	planar	0.0158
nanluoguxiang	0.0012	Stadium	0.0179	lighting	0.012	sky	0.0132
OMA	0.0012	Chengde	0.0175	lion	0.0116	shop	0.0110
USER	**PROB.**	**USER**	**PROB.**	**USER**	**PROB.**	**USER**	**PROB.**
J0N6	0.1107	flomuc79	0.3065	neonbubble	0.3231	*ken*	0.8639
Miki Badt	0.0819	NursiePoo	0.0842	sndgrss	0.0492	Bowen LU	0.0072
Yuwei*	0.0345	a. j. miller	0.0525	BootsinOven	0.0303	ChrisKSdub	0.0063
kirkario	0.0225	A. Folly	0.0518	Eights	0.0205	Wei	0.0061
nancysagar	0.0196	Lemmo2009	0.0376	E. Oliver	0.0192	Egyptian Mau	0.0055
yiduiqie	0.0195	Schwarzers	0.0329	Y. Zhang	0.0179	Myarmy	0.0048
T. Nylund	0.0179	ochurchill	0.0322	ada	0.0166	GraemeNicol	0.0042
flomuc79	0.0170	A. Lai	0.0252	GraemeNicol	0.0159	luca64bj	0.0029
Choollus	0.0146	Srikanta	0.0147	fpsanders	0.0146	rosoon	0.0016
Schwarzers	0.0138	Beautiful	0.0118	trackcycling	0.0127	*DigiPub*	0.0012
choubb	0.0132	C. Kealy	0.0092	m.tomic	0.0107	m. braun	0.0012
T. Simkin	0.0123	FD	0.0085	R. Kendall	0.0101	le niners	0.0010

Note: We highlight the occurrence of the top two users of Fig.3 with red color in this table, to point out the association of this table with Fig.3.

Table II provides the intuitional explanations of contents or theme for these 8 picked latent topics.

TABLE II
INTUITIONAL EXPLANATIONS FOR THE PICKED PREFERENCE TOPICS

Topic index	Intuitional explanations
10	Chinese cities, Location
27	Animal, insect species and its color
48	Fine food, delicacies
70	Nature scenes and weather condition
80	Local parks in Beijing
84	Local palaces in Beijing
86	Athletics, Sport Events
95	Camera brands, varieties and model

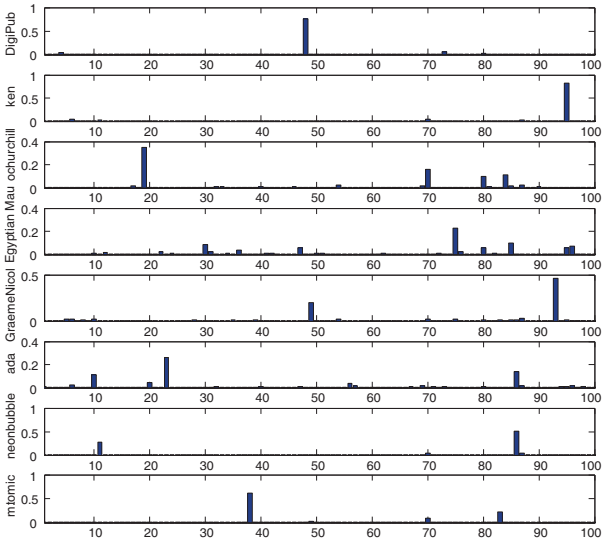

Fig. 3. Users' preference-topic distributions. In each subplot, x-axis represents the topic index and y-index represents the probability of a user's interest on different topics.

It can be observed that most of the tags distribution conditioned on topics make sense. It is also interesting to observe that the tag "Honeymoon" appears in topic 86 (sports), which seems inappropriate and is a outlier. A careful examination of the dataset[2] however reveals that a rather significant number of couples in fact spent their honeymoon in the Bejing 2008 Olympic Games and took pictures near the sport facilities . Fig. 3 also shows that "neonbubble" and "ada" share the common interest in sports, but in addition to sports, "ada" also like to spend his spare time with animals and insects. And "Egyptian Mau" have relative broader preferences.

C. Discovered Preference-Topic Distributions

In a more intuitive way, Fig. 3 reorganizes the inferred $p(t|a)$ (i.e., θ_a) of several representative users picked out from Table I. As we can observed, "ken" is most likely a digital camera fan, with the probability $p(t = 95|a = \text{"}ken\text{"}) = 0.8639$ favoring topic 95, and like taking landscape photos,

[2]http://www.flickr.com/photos/neonbubble/3081485579/

(a) The tag cloud on ken's profile

(b) Part of DigiPub's tag cloud

Fig. 4. Two snapshots from Flickr site to confirm our preference topic discovering method. And we use rectangular frame with different colors to highlight the tags obviously supporting our experimental result.

with the probability $p(t = 70|a = \text{“}ken\text{”}) = 0.02443$. Fig. 4(a) shows the snapshot of ken's frequently used tags on his personal page, the highlighted tag about camera model and other photographic terms obviously support our experiment result. however, "DigiPub" is probably an epicures, a part of DigiPub's tags in Fig. 4(b) indicates his strong preference of delicacies and relatively weaker preference for insect. The 5 Chinese traditional photos attached to "chinatown" captured from the URL[3], means the intent behind this tag is indeed relevant to food too.

D. Potential Applications

The latent random variable *preference topic*, in this study, is used to model users' preferences discovered from the tags annotated on the photos. But in general, the user's preference topic distributions can directly express their interests of diverse aspect of the emergent semantic structure. This fact could be widely used in the current thriving social media communities.

One application of our Preference-Topic model is the user segmentation, by clustering user according to the preference distributions. Because the user preference topic distribution vector is distributed over the K-dimensional unit hypersphere and with equal magnitude, we employ hierarchical clustering algorithm incorporating cosine similarity as the proximity measure to address this application problem. The clustering result on the basis of the last experiment step is visualized in Fig.5. It is obvious that the users residing in the middle of the matrix have extensive preferences, but the users' preferences distribution on the left and right end are very sparse. And the leftmost 3 clusters take on a pretty cohesion.

Furthermore, There is another potential application scenario of multimedia retrieval in folksonomy supported system. When a search request comes, we could reformulate the query by incorporating the latent preferred topics learned by our model as the context same as the intent or geographical constraints. The results recalled by the original query requirements are filtered through comparing media topics to user's preference topics. More broadly speaking, our model can be tuned to serve as a core component in personalized community, recommender system, and multimedia search engine, treating the users' preferences as a kind of important metadata in user profile.

[3]http://www.flickr.com/photos/pcfannet/tags/chinatown/

V. Conclusion and Future Works

In this paper we present the Preference-Topic model for the inference of user preference in online communities with folksonomy. More specifically, by introducing a latent preference variable, we model the dependency structure of the users' preferences, photos and tags in a generative manner. Furthermore, we provide the proposed model with a theoretical interpretation and a mathematical estimation technique. Dataset from Flickr is used to verify our model and inference algorithm. Our experimental results illustrate that the proposed model can effectively assist us to discover the aspects mentioned above. Finally, we give two application scenarios of user segmentation and query refinement, to confirm the significant application value and diversity of our preference discovering model.

In our future work, we intend to improve our model to measure the similarity between the photo's topic and user's topic, so that to deal with the web media search refinement and recommendation considering user's preference. Furthermore, we will incorporate the rating and more richer implicit feedback data to modify our model to a supervised model to perform social media search and recommendation. Last but not least, as we note that our algorithm performance is sensitive to the model's Dirichlet prior hyper-parameters α and β and topic number. This fact make it desirable to introduce a unified model selection schema, for example, comparison of

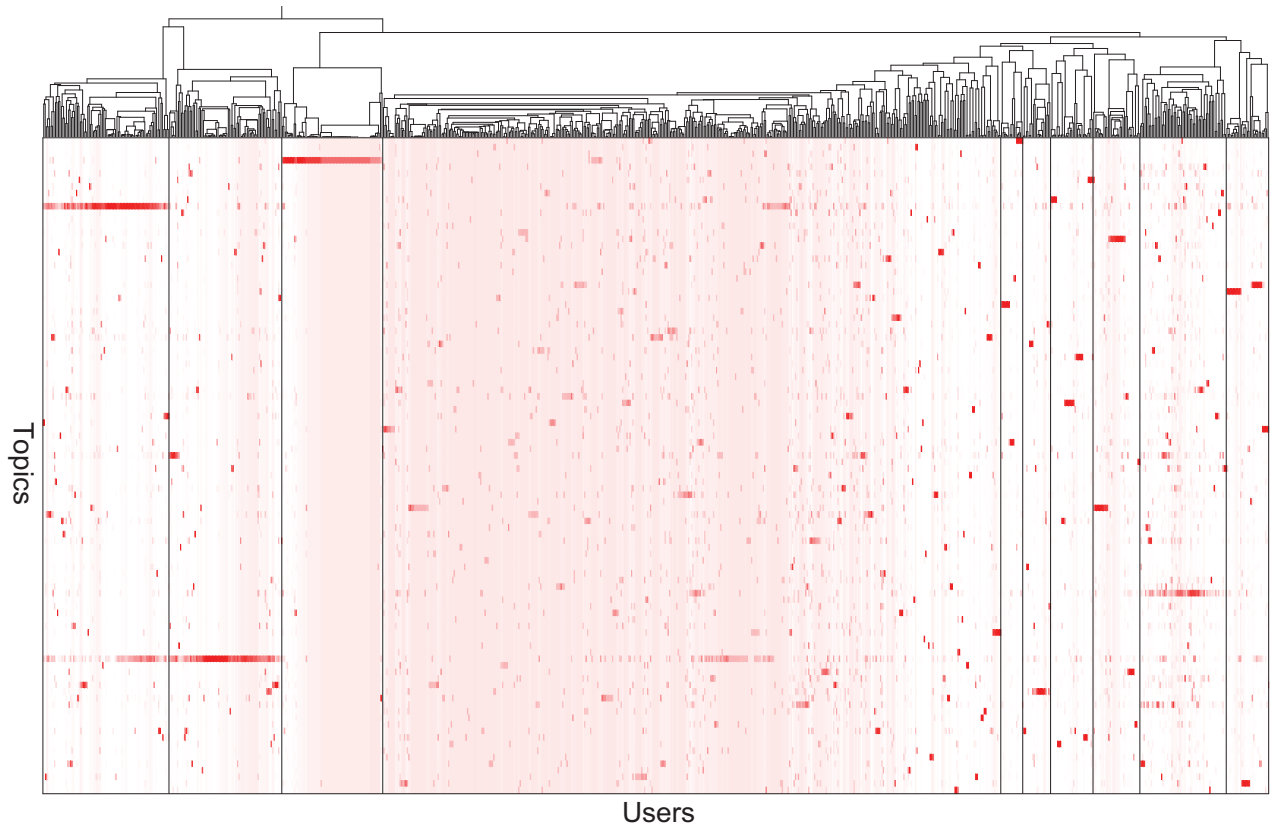

Fig. 5. The matrix visualization of user clustering result with 10 clusters, in term of the discovered user preference topic distribution. The rows are indexed by the 100 latent topics, the columns denote the users preference probability distribution, and the redder cell represents the higher probability value. The hierarchical agglomerative tree is on the top of the matrix. The vertical black lines divide the users into 10 clusters.

Bayes factors, the log-likelihood probability[4], perplexity in language model, or developing a nonparametric model[13], for our proposed user preference discovery approach.

VI. Acknowledge

This work was supported partly by National Natural Science Foundation of China (No. 61103031), partly by China 863 program (No. 2012AA011203), partly by the State Key Lab for Software Development Environment (SKLSDE-2013ZX-16), partly by the Fundamental Research Funds for the Central Universities (No. YWF-12-RHRS-016).

References

[1] D. M. Blei, A. Y. Ng, M. I. Jordan, and J. Lafferty. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:1107–1135, 2003.

[2] K. Farrahi and D. Gatica-Perez. Discovering routines from large-scale human locations using probabilistic topic models. *ACM Trans. Intell. Syst. Technol.*, 2(1):3:1–3:27, Jan. 2011.

[3] J. Fürnkranz and E. Hüllermeier. Preference learning: An introduction. In *Preference Learning*, pages 1–17. Springer, 2011.

[4] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

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

[6] N. Indurkhya and F. J. Damerau. *Handbook of Natural Language Processing*. Chapman & Hall/CRC, 2nd edition, 2010.

[7] R. Krestel and P. Fankhauser. Personalized topic-based tag recommendation. *Neurocomputing*, 76(1):61 – 70, 2012.

[8] Y. Lu, C. Zhai, and N. Sundaresan. Rated aspect summarization of short comments. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 131–140, New York, NY, USA, 2009. ACM.

[9] A. McCallum, X. Wang, and A. Corrada-Emmanuel. Topic and role discovery in social networks with experiments on enron and academic email. *J. Artif. Int. Res.*, 30(1):249–272, Oct. 2007.

[10] T. Minka and J. Lafferty. Expectation-propagation for the generative aspect model. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, UAI'02, pages 352–359, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.

[11] M. Rosen-Zvi, C. Chemudugunta, T. Griffiths, P. Smyth, and M. Steyvers. Learning author-topic models from text corpora. *ACM Trans. Inf. Syst.*, 28(1):4:1–4:38, Jan. 2010.

[12] A. Sun and S. S. Bhowmick. Image tag clarity: in search of visual-representative tags for social images. In *Proceedings of the first SIGMM workshop on Social media*, WSM '09, pages 19–26, New York, NY, USA, 2009. ACM.

[13] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[14] H. Xu, J. Wang, X.-S. Hua, and S. Li. Tag refinement by regularized lda. In *Proceedings of the 17th ACM international conference on Multimedia*, MM '09, pages 573–576, New York, NY, USA, 2009. ACM.

[15] J. Zeng, W. K. Cheung, and J. Liu. Learning topic models by belief propagation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(5):1121–1134, 2013.