10-2012

# A Probabilistic Graphical Model for Topic and Preference Discovery on Social Media

Lu LIU
*Capital Medical University*

Feida ZHU
*Singapore Management University*, fdzhu@smu.edu.sg

Lei ZHANG
*Microsoft Research Asia*

Shiqiang YANG
*Tsinghua University*

**DOI:** https://doi.org/10.1016/j.neucom.2011.05.039

Citation

LIU, Lu; ZHU, Feida; ZHANG, Lei; and YANG, Shiqiang. A Probabilistic Graphical Model for Topic and Preference Discovery on Social Media. (2012). *Neurocomputing*. 95, 78-88. Research Collection School Of Information Systems.
**Available at:** https://ink.library.smu.edu.sg/sis_research/3209

Published in Neurocomputing, Volume 95, 15 October 2012, Pages 78-88.
https://doi.org/10.1016/j.neucom.2011.05.039

# A probabilistic graphical model for topic and preference discovery on social media

Lu Liu [a,*], Feida Zhu [d], Lei Zhang [c], Shiqiang Yang [b]

[a] Capital Medical University, China
[b] Tsinghua University, China
[c] Microsoft Research Asia, China
[d] Singapore Management University, Singapore

*Keywords:*
Social media mining
Topic model
Preference discovery

## ABSTRACT

Many web applications today thrive on offering services for large-scale multimedia data, e.g., Flickr for photos and YouTube for videos. However, these data, while rich in content, are usually sparse in textual descriptive information. For example, a video clip is often associated with only a few tags. Moreover, the textual descriptions are often overly specific to the video content. Such characteristics make it very challenging to discover topics at a satisfactory granularity on this kind of data. In this paper, we propose a generative probabilistic model named Preference-Topic Model (PTM) to introduce the dimension of user preferences to enhance the insufficient textual information. PTM is a unified framework to combine the tasks of user preference discovery and document topic mining together. Through modeling user-document interactions, PTM cannot only discover topics and preferences simultaneously, but also enable them to inform and benefit each other in a unified framework. As a result, PTM can extract better topics and preferences from sparse data. The experimental results on real-life video application data show that PTM is superior to LDA in discovering informative topics and preferences in terms of clustering-based evaluations. Furthermore, the experimental results on DBLP data demonstrate that PTM is a general model which can be applied to other kinds of user–document interactions.

## 1. Introduction

Nowadays, many web applications, e.g., Flickr and YouTube, thrive on offering services for the interactions between users and media content and produce a new type of multimedia content termed as ‘‘social media’’. The interactions between users and social media include various user behaviors, e.g., publishing, accessing, annotating images or videos as shown in Fig. 1. Essentially these user behaviors are closely correlated with social media topics as well as user interests. And social media topic discovery and user preference modeling are important problems which have attracted considerable research interests.

Interactions have supplied social media with user-contributed contextual information, such as tags, titles, etc., which can be used to infer social media topics and web user interests. However, most of the textual annotations are usually sparse and overly specific. For example, each online video clip is often associated with only a few tags. We counted the number of tags of 157,520 videos clips on YouKu,[1] one of the largest video sharing websites in China, and found that the average number of tags for a video is only 3.77 with the maximum being 29. The distribution of the number of tags for the video set shown in Fig. 2 illustrates that more than 50% videos have fewer than four tags. Clearly, the textual descriptions of web videos are very short and sparse. On the other hand, it is observed that users tend to label videos with specific tags to indicate the video content. Table 1 shows five videos and their tags from YouTube. It is easy to see that their tags can be so specific that videos, even belonging to the same topic (e.g., news, soccer in Table 1), may have completely non-overlapping tags.

Such sparse and overly specific textual data pose a tough challenge for social media topic mining. For example, it is very difficult to assign video 4 and video 5 in Table 1 to the same category based only on their tags. Simple vector space model based approaches cannot well handle this problem due to the textual mismatch problem. In recent years, many probabilistic topic models have been proposed for text mining tasks. For example, PLSA [1] and LDA [2] are capable of discovering latent

* Corresponding author.
*E-mail address:* liulu26@gmail.com (L. Liu).

[1] http://www.youku.com

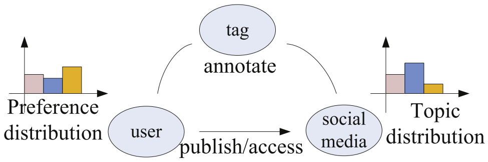

**Fig. 1.** The interactions between social media and web users.

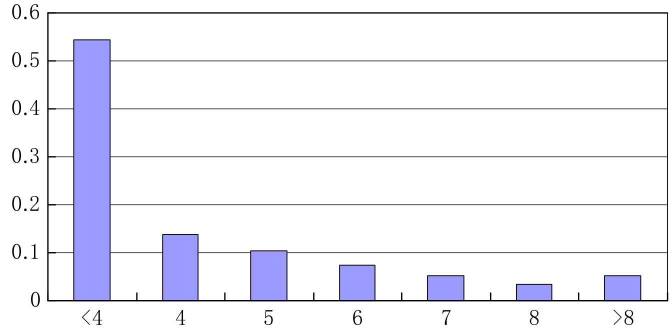

**Fig. 2.** The proportion of videos vs. the number of tags (statistics of 157,520 videos on YouKu, a video sharing website in China).

**Table 1**
Examples of tags for videos on YouTube.

| News | |
|---|---|
| Video 1 | Israel palestinians threatens retaliation gaza |
| Video 2 | BBC News Heavy snow in weather forecast |
| Video 3 | winter storm 700 000 still ky |
| Soccer | |
| Video 4 | Cagliari Calcio Juventus Match Partita Hurucan |
| Video 5 | A amazing goal fabio udinese Napoli Italy |

topics of documents. Besides textual data, some researchers have incorporated other information into topic modeling, such as time [3], geographic location [4], authorship [5–7], and network structure [8,9]. However, none of these models study the problem of topic modeling on sparse and specific textual descriptions of social media. It is generally ineffective for topic models to discover latent topics from textual descriptions of social media due to the lack of redundant textual information.

Fortunately, besides the textual descriptions, the interactions between users and social media provide abundant user action logs, such as viewing videos (e.g., YouTube), reading blogs (e.g., MSN blog), sharing pictures (e.g., Flickr), buying commodities (e.g., Amazon). These data bring a large amount of supplementary information, which can enhance the insufficient textual descriptions. For example, users who are interested in soccer are likely to view both video 4 and video 5 in Table 1. Through this piece of user interaction information, even videos without common tags can be identified as similar ones which belong to the same or related topics. It is therefore evident that user interaction information can be practically exploited to improve the performance of topic mining tasks on social media data.

In this paper, we study the challenging problem of discovering topics and user preferences on social media data simultaneously. We propose a unified generative probabilistic model named Preference-Topic model (PTM) to combine the user logs and textual information together and leverage the sufficient user log information to introduce a dimension of preferences in the model. Preferences, which reflect user interest patterns, are modeled as higher-level constraints over topics to supplement the insufficient textual descriptive information. As shown in Fig. 1, user behavior is generated based on their preference distributions while document content is drawn according to their topic distributions. Furthermore, these two parts are connected by the interactions between users and documents. Thus, PTM is a unified framework which combines the tasks of user interest modeling and document topic discovery together. PTM cannot only discover topics and preferences simultaneously, but also enable them to benefit each other by modeling them in the unified framework.

In the experiments, we apply PTM to two types of user–document interaction data including the typical social media data from YouKu. First the results are illustrated to demonstrate that PTM can discover meaningful topics and preferences. Then we evaluate the quality of topics and preferences and compare them with LDA in terms of video and user clustering. The comparison results demonstrate that PTM is superior to LDA in discovering more informative topics and preferences. In particular, it proves that topics and preferences in a unified framework indeed benefit each other in the following ways:

- Since the user–interaction information enhances the sparse textual information, the unified framework helps to discover more informative topics, as demonstrated by the better performance of video clustering in the experiments.
- As a preference, which is represented as a multinomial distribution over topics, is a higher-level description over topics, the related topics are likely to be generated conditioned on the same preference. Therefore preferences enhance the sparse textual information by bringing the related topics under a common theme at a higher abstract level. For example, as the textual descriptions are sparse and specific, video 2 and video 3 in Table 1 may be assigned to different but related topics. However, preferences could then reveal the connections between the topics so that the two videos would be classified to the same cluster at a higher level.
- For large-scale data, it is common that users with similar interests would interact with disjoint sets of social media. In this case, topics can help reveal their similar interests. For example, suppose two users view different sets of videos on the common topic “soccer”. Then it is easy to get that these two users have similar interests based on these videos’ topics. In other words, topics are also helpful for learning better preferences, as illustrated by the better quality of inferred user clusters in the experiments.

The remainder of this paper is organized as follows. We define the problem in Section 2. Preference-Topic model is described in Section 3 and the parameter estimation results by variational inference are given in Section 4. The experimental result illustration and the performance comparisons are presented in Section 5. In Section 6, we discuss related work. The paper is concluded in Section 7.

## 2. Problem formulation

In this section, we present several necessary definitions for the tasks of mining user interests and document topics.

**Definition 1** (*Document*). A document $d$ is defined as a sequence of $K$ words, i.e., $d = (w_1, \ldots, w_K)$, where each word is chosen from a vocabulary of size $V$.

In this paper, we define a document as an abstract concept to represent any type of social media, e.g., a blog, an image, a video, etc.

**Definition 2** (*Interaction*). An action which is produced by a user on a document is defined as an interaction $c$ between the user and the document.

There are various kinds of interactions, such as viewing videos, reading blogs, sharing pictures, buying commodities, which can be obtained from user logs of social media.

**Definition 3** (*Document Topic Modeling*). A topic is defined as a multinomial distribution over words $p(w|\beta)$. Each document is represented as a mixture of topics.

Therefore, words with the highest probability in the distribution would suggest the semantics of the topic. Our goal is to discover the topics on social media.

**Definition 4** (*User Preference Mining*). A preference is defined as a pattern which reflects user interests. Each user is represented as a mixture of preferences.

The other goal of this paper is to mine the preferences and discover use interests. There are various ways to define the distributions of preferences. E.g., if LDA is employed to discover user interests, a preference is defined as a multinomial distribution over documents. In this paper, we define a preference as a high-level description, which is represented as a multinomial distribution over topics as discussed in Section 3.

## 3. Preference-Topic model

In this section, we introduce Preference-Topic model, which combines the logs and textual information together to mine user interests and document topics simultaneously. First, we give two general assumptions to build the model.

**Assumption 1.** User behaviors are generated based on user interests in social media networks.

In reality, the driving factors of user behaviors are complex and subtle. For example, we may view a web video based on friends' recommendation or simply by pure chance. In this paper, our goal is to discover the major factor of user behaviors, i.e., user interests. Therefore we use Assumption 1 as the foundation of our model. More factors, e.g., user influence, have been explored to model user behaviors in our follow-up work [10].

As each user's interest is represented as a mixture of preferences in our model, this assumption indicates that user behaviors, e.g., viewing videos or photos, are generated conditioned on users' preference distributions.

**Assumption 2.** When a user interacts with a document, he/she must be attracted by one topic of the document. In other words, the interaction happens when one topic of the document matches user interest.

This assumption supplements and enhances the first assumption. As our Preference-Topic model represents each document's content as a mixture of topics, this assumption further stipulates that a particular user action happens based on not only the user's preference distribution but also the corresponding document's topic distribution.

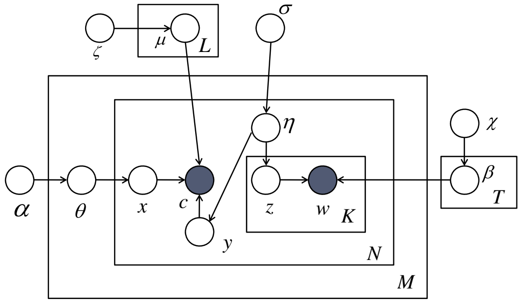

**Fig. 3.** Graphical model representation of Preference-Topic model. The boxes are "plates" representing replicates. The outer plate represents users, the middle plate represents interactions related to a user and the corresponding documents while the inner plate represents the repeated choice of topics and words within a document.

### 3.1. Generative process

Based on the assumptions above, we propose a probabilistic graphical model named Preference-Topic model as shown in Fig. 3. The boxes in the figure represent replicates. The outer plate represents users. The middle plate represents interactions related to a user and the corresponding documents while the inner plate represents the repeated choice of topics and words within a document.

According to different levels of replicates, the generative process is hierarchical as illustrated in Algorithm 1:

- On the first level, for each user, his/her preference distribution is sampled, which is indicated by the parameter $\theta$.
- On the second level, for each interaction of this user, the corresponding document's topic distribution $\eta$ is sampled. And based on both of the document topic distribution $\eta$ and user preference distribution $\theta$, a topic $y$ and a preference $x$ are generated and then the interaction $c$ is generated.
- On the third level, based on document topic distribution $\eta$, each document's topics $z$ and words $w$ are generated.

The variable descriptions are shown in Table 2. In particular, the two gray nodes in Fig. 3 represent the observed variables of document word and user interaction respectively. Other nodes denote the hidden variables and parameters. In summary, w.r.t. hierarchical generative process, the corpus-level parameters $\alpha,\beta,\mu,\sigma$ are assumed to be sampled once in the process of generating a corpus; the user-level variable $\theta$ is sampled once per user; the document-level variables $c,x,y,\eta$ are sampled once per document; while the word-level variables $z$ and $w$ are sampled once per word.

Based on the generative process, it is easy to get the posterior probability to generate all the interactions of a user as below:

$$p(\mathbf{x},\mathbf{y},\mathbf{Z},\theta,\boldsymbol{\eta},\mathbf{c},\mathbf{W}|\alpha,\beta,\mu,\sigma)$$
$$= p(\theta|\alpha)\prod_{n=1}^{N} p(x_n|\theta)p(y_n|\eta_n)p(c_n|x_n,y_n,\mu)p(\eta_n|\sigma)\prod_{k=1}^{K} p(z_{nk}|\eta_n)p(w_{nk}|z_{nk},\beta) \tag{1}$$

Integrating or summing all the hidden variables $\mathbf{x},\mathbf{y},\mathbf{Z},\theta,\boldsymbol{\eta}$, we obtain the marginal distribution for one user

$$p(\mathbf{c},\mathbf{W}|\alpha,\beta,\mu,\sigma) = \int_{\theta}\sum_{\mathbf{x}}\sum_{\mathbf{y}}\sum_{\mathbf{Z}}\int_{\boldsymbol{\eta}} p(\mathbf{x},\mathbf{y},\mathbf{Z},\theta,\boldsymbol{\eta},\mathbf{c},\mathbf{W}|\alpha,\beta,\mu,\sigma) \tag{2}$$

Eq. (2) is used as the objective function to be maximized for the model estimation as each user is assumed to be independent in the model.

**Table 2**
Variable descriptions.

| | |
|---|---|
| *Observed variables* | |
| $w$ | A word from a vocabulary indexed by $\{1,\ldots,V\}$. A document is a sequence of $K$ words denoted by $\mathbf{w}=(w_1,w_2,\ldots,w_K)$. And all the documents related to a user are denoted by $\mathbf{W}=(\mathbf{w}_1,\mathbf{w}_2,\ldots,\mathbf{w}_N)$ |
| $c$ | An interaction between a user and a document. A user is represented by a sequence of $N$ related interactions $\mathbf{c}=(c_1,c_2,\ldots,c_N)$ |
| $N$ | The number of documents related to a user |
| $M$ | The number of users |
| $K$ | the number of words in a document |
| *Hidden variables* | |
| $z_{nk}$ | The topic which generates the word $k$ in document $n$. Let $\mathbf{Z}=(z_{1,1},\ldots,z_{1,K};z_{2,1},\ldots,z_{2,K};\ldots;z_{N,1},\ldots,z_{N,K})$ be all the latent topics of the documents that a user interacts with |
| $x_n$ | The preference which generates the interaction $c_n$. Let $\mathbf{x}=(x_1,x_2,\ldots,x_N)$ be the set of latent preferences that generate all the interactions related to a user |
| $y_n$ | The topic of document $n$ that the user is interested in. Let $\mathbf{y}=(y_1,y_2,\ldots,y_N)$ be the set of latent topics that a user is interested in |
| $\theta$ | A $L$-dimensional Dirichlet random variable that represents the multinomial distribution of a user over preferences |
| $\eta_n$ | A $T$-dimensional Dirichlet random variable that represents the multinomial distribution of the document over topics. Suppose $\boldsymbol{\eta}=(\eta_1,\eta_2,\ldots,\eta_N)$ |
| *Variational variables* | |
| $\gamma$ | A Dirichlet parameter to estimate $\theta$ |
| $\phi_n$ | A multinomial parameter to estimate each user's preference distribution |
| $\delta_n$ | A multinomial parameter to estimate the topic distribution of each interaction |
| $\omega_n$ | A Dirichlet parameter to estimate $\eta_n$ |
| $\lambda_{nk}$ | A multinomial parameter to estimate each document's topic distribution |
| *Parameters* | |
| $\alpha$ | The Dirichlet prior of user multinomial distributions over preferences |
| $\sigma$ | The Dirichlet prior of document multinomial distributions over topics |
| $\beta$ | The conditional multinomial distributions of topics over words |
| $\mu$ | The conditional multinomial distributions of preferences over topics |
| *Input parameters* | |
| $T$ | The number of topics |
| $L$ | The number of preferences |

### 3.2. Comparison with LDA

LDA model [2] is a classic probabilistic model which can be employed to discover document topics. The generative process works as: for each document, a multinomial distribution is generated from Dirichlet prior, based on which the topics and the words in the document are generated.

If each user is treated as a document and each interaction is treated as a word, then LDA can also be employed to discover user interest patterns on user logs [2]. However, it is hard to describe the semantics of patterns as no textual information is utilized.

Compared with LDA, PTM has added a higher level generation to capture the latent user preference distributions. PTM represents each document as a mixture of topics and each user as a mixture of preferences. And a preference is represented as a higher-level description over topics. PTM focuses on modeling the generation of interactions between users and documents and assumes that each interaction is generated when user preference matches document topic which is controlled by the parameter $\mu$.

Notice that when a user views a document, a new copy of the document is generated in the model. The reason for this way of modeling is that a document is deemed different when viewed by different users. Thus when PTM generates a document, it actually generates a pseudo-document, which reflects user interests. Therefore when dealing with each user, PTM needs to generate each document corresponding to the user.

As the number of documents related to each user $N$ is small which can be deemed as a constant, the computation cost and runtime of PTM is proportional to the number of users $M$. In contrast, the computation cost of LDA is proportional to the number of documents. Both models can use parallel computing to increase the efficiency of the model estimation process.

**Algorithm 1.** Probabilistic generative process.

```
foreach user, sample $\theta \sim Dir(\alpha)$ do
    foreach interaction $c_n$ related to the user do
        for the corresponding document, sample $\eta_n \sim \mathrm{Dir}(\sigma)$
        foreach word $k$ do
            sample a topic $z_{nk} \sim$ Multinomial $(\eta_n)$;
            sample the word $w_{nk} \sim p(w|z_{nk},\beta)$;
        end
        sample a preference $x_n \sim$ Multinomial $(\theta)$;
        sample a topic $y_n \sim$ Multinomial $(\eta_n)$;
        sample the interaction $c_n \sim p(c_n|x_n,y_n,\mu)$;
    end
end
```

## 4. Model estimation

### 4.1. Variational inference

As the likelihood of the data Eq. (2) is intractable due to the coupling between hidden variables, the posterior distributions cannot be exactly inferred [2]. A variety of algorithms have been used to solve the problem, including variational approximation [2] and Gibbs sampling [11]. In this paper, we employ the variational inference method to approximate the posterior distributions of latent variables. The basic idea of variational inference is first to get a lower bound of the likelihood and then to maximize the likelihood via maximizing the lower bound.

By using Jensen's inequality, an adjustable lower bound of the log likelihood can be obtained as [12]

$$\log p(\mathbf{c},\mathbf{W}|\alpha,\beta,\mu,\sigma)$$
$$\geq \int_\theta \sum_{\mathbf{x}} \sum_{\mathbf{y}} \sum_{\mathbf{Z}} \int_{\boldsymbol{\eta}} q(\mathbf{x},\mathbf{y},\mathbf{Z},\theta,\boldsymbol{\eta}) \log p(\mathbf{x},\mathbf{y},\mathbf{Z},\theta,\boldsymbol{\eta},\mathbf{c},\mathbf{W}|\alpha,\beta,\mu,\sigma)$$

$$-\int_{\theta}\sum_{\mathbf{x}}\sum_{\mathbf{y}}\sum_{\mathbf{Z}}\int_{\boldsymbol{\eta}}\log q(\mathbf{x},\mathbf{y},\mathbf{Z},\theta,\boldsymbol{\eta}) \tag{3}$$

where $q(\mathbf{x},\mathbf{y},\mathbf{Z},\theta,\boldsymbol{\eta}|\gamma,\phi,\delta,\omega,\lambda)$ could be any arbitrary variational distribution with variational parameters $\gamma$, $\phi$, $\delta$, $\omega$, $\lambda$.

Let $L(\gamma,\phi,\delta,\omega,\lambda:\mathbf{x},\mathbf{y},\mathbf{Z},\theta,\boldsymbol{\eta})$ denotes the right-hand side of the above inequation, which is also the lower bound of the log likelihood. It can be easily verified that the distance between the lower bound and the log likelihood is the Kullback–Leibler (KL) divergence between the variational distribution and the true posterior distribution, i.e.

$$\begin{aligned}
&\log p(\mathbf{c},\mathbf{W}|\alpha,\beta,\mu,\sigma)\\
&\quad = L(\gamma,\phi,\delta,\omega,\lambda:\mathbf{x},\mathbf{y},\mathbf{Z},\theta,\boldsymbol{\eta})\\
&\qquad + KL(q(\mathbf{x},\mathbf{y},\mathbf{Z},\theta,\boldsymbol{\eta})\|p(\mathbf{x},\mathbf{y},\mathbf{Z},\theta,\boldsymbol{\eta},\mathbf{c},\mathbf{W}|\alpha,\beta,\mu,\sigma))
\end{aligned} \tag{4}$$

Thus, maximizing the lower bound equals to minimizing the KL-divergence.

In order to solve the intractable problem, we define the following variational distribution on the latent variables:

$$\begin{aligned}
&q(\mathbf{x},\mathbf{y},\mathbf{Z},\theta,\boldsymbol{\eta}|\gamma,\phi,\delta,\omega,\lambda)\\
&\quad = q(\theta|\gamma)\prod_{n=1}^{N} q(x_n|\phi_n)q(y_n|\delta_n)q(\eta_n|\omega_n)\prod_{k=1}^{K} q(z_{nk}|\lambda_{nk})
\end{aligned} \tag{5}$$

where Dirichlet parameters $\gamma,\omega_n$ and multinomial parameters $\phi_n,\delta_n,\lambda_{nk}$ are variational parameters. The meaning of these variational parameters are also described in Table 2.

By maximizing the lower bound, we can estimate the variational parameters by the following equations:

$$\gamma_i = \alpha_i + \sum_{n=1}^{N}\phi_{ni} \tag{6}$$

$$\phi_{ni} \propto \exp\left(\Psi(\gamma_i) - \Psi\left(\sum_{j=1}^{L}\gamma_j\right) + \sum_{j=1}^{T}\delta_{nj}\log(\mu_{ij})\right) \tag{7}$$

$$\delta_{ni} \propto \exp\left(\sum_{j=1}^{L}\phi_{nj}\log(\mu_{ji}) + \Psi(\omega_{ni}) - \Psi\left(\sum_{j=1}^{T}\omega_{nj}\right)\right) \tag{8}$$

$$\omega_{ni} = \sigma_i + \sum_{k=1}^{K}\lambda_{nki} + \delta_{ni} \tag{9}$$

$$\lambda_{nki} \propto \beta_{iw}\exp\left(\Psi(\omega_{ni}) - \Psi\left(\sum_{j=1}^{T}\omega_{nj}\right)\right) \tag{10}$$

where $\Psi$ is the first derivative of the $\log\Gamma$ function which is computable via Taylor approximations.

Eq. (7) shows that $\phi$ is determined by two parts: $\Psi(\gamma_i) - \Psi(\sum_{j=1}^{L}\gamma_j)$ and $\sum_{j=1}^{T}\delta_{nj}\log(\mu_{ij})$. It indicates that when an interaction between a user and a document happens, the selected preference probability is determined not only by the multinomial distribution of the user preferences, but also by the product of transition probability from preferences to topics and the selected topic probability of the document. Eqs. (8) and (9) have similar explanations. Thus it indicates that in principle the discovered preferences and topics indeed impact each other in the model.

The variational inference procedure to estimate the variational parameters is summarized as Algorithm 2.

**Algorithm 2.** Variational inference procedure for variational parameter estimation.

```
Initialize;
for $i=1$ to $L$, $n=1$ to $N$ do
| $\phi_{ni}^0 = 1/L, \gamma_i^0 = \alpha + N/L$;
end
for $i=1$ to $T$, $n=1$ to $N, k=1$ to $K$ do
| $\delta_{ni}^0 = 1/T, \omega_{ni}^0 = \sigma_i + (K+1)/T, \lambda_{nki}^0 = 1/T$;
end
Repeat;
for $n=1$ to $N$ do
    for $k=1$ to $K$ do
        for $i=1$ to $T$ do
            update $\lambda_{nki} \propto \beta_{iw}\exp(\Psi(\omega_{ni}) - \Psi(\sum_{j=1}^{T}\omega_{nj}))$
        end
        normalize $\lambda_{nki}$ to sum to 1;
    end
    for $i=1$ to $T$ do
        update $\omega_{ni} = \sigma_i + \sum_{k=1}^{K}\lambda_{nki} + \delta_{ni}$
    end
    for $i=1$ to $T$ do
        update $\delta_{ni} \propto \exp(\sum_{j=1}^{L}\phi_{nj}\log(\mu_{ji}) + \Psi(\omega_{ni}) - \Psi(\sum_{j=1}^{T}\omega_{nj}))$
    end
    normalize $\delta_{ni}$ to sum to 1;
    for $i=1$ to $L$ do
        update $\phi_{ni} \propto \exp(\Psi(\gamma_i) - \Psi(\sum_{j=1}^{L}\gamma_j) + \sum_{j=1}^{T}\delta_{nj}\log(\mu_{ij}))$
    end
    normalize $\phi_{ni}$ to sum to 1
end
for $i=1$ to $L$ do
    update $\gamma_i = \alpha_i + \sum_{n=1}^{N}\phi_{ni}$
end
```

### 4.2. Parameter estimation

As variational inference provides us with a lower bound, we use it as a surrogate for the marginal likelihood and employ an empirical Bayes method via an alternating variational EM procedure to estimate the parameters.

E-step for each user, find the optimizing values of the variational parameters as described in Section 4.1.

M-step maximize the resulting lower bound on the log likelihood in Eq. (11) with respect to the model parameters $\alpha$, $\beta$, $\mu$, $\sigma$

$$L(\alpha,\beta,\mu,\sigma) = \sum_{m=1}^{M}\log p(\mathbf{c_m},\mathbf{W_m}|\alpha,\beta,\mu,\sigma) \tag{11}$$

The two steps are repeated until the lower bound converges.

The M-step update for the conditional multinomial parameters $\mu$, $\beta$ can be estimated analytically

$$\mu_{ij} \propto \sum_{m=1}^{M}\sum_{n=1}^{N}\phi_{mni}\delta_{mnj} \tag{12}$$

$$\beta_{ij} \propto \sum_{m=1}^{M}\sum_{n=1}^{N}\sum_{k=1}^{K}\lambda_{mnki}w_{mnk}^{j} \tag{13}$$

The Dirichlet parameter $\alpha,\sigma$ can be estimated by using an efficient Newton–Raphson method [2] which is not provided in detail in the paper.

## 5. Experiments

### 5.1. Experimental setup

PTM is general to any kind of user–document interaction data in social media networks, such as reading blogs, buying commodities, writing papers, etc. In this paper, we collect two types of interaction data to evaluate the performance of the proposed approach.

- *YouKu media data*: YouKu, the counterpart of Youtube in China, is a popular video sharing website. We acquired a data set which includes user access logs from 1–15 November in 2006 as well as the corresponding video titles and tags directly from YouKu as we have cooperation.
- *DBLP publication data*: We also collect the papers on four conferences: SIGIR, WWW, KDD, NIPS and the corresponding authors from DBLP data set.

The algorithms are implemented in C++ and run on an Intel Core 2 T7200 and a processor with 2GB DDR2 RAM. Like other topic models, the training time of the model also depends on the parameters, e.g., the number of preferences and topics, the iterations, etc., and it requires several hours. We evaluate our method on the following three aspects:

- *Topic and preference illustration*: We first illustrate the extracted topics and preferences to demonstrate that PTM can get meaningful results on the user–document interaction data sets.
- *Video and user clustering performance evaluation*: Then, we evaluate the quality of topics and preferences in terms of video and user clustering performance. We compare the results of PTM with LDA to show that PTM can extract better topics and preferences than LDA.
- *Perplexity test*: Thirdly, we compare the perplexity of PTM with LDA to illustrate that PTM has better generalization performance than LDA.

**Table 3**
Statistics of YouKu media data.

| Data set | # users | # docs | # words | d/u | w/d |
|---|---|---|---|---|---|
| Illustration | 16,775 | 9100 | 12,584 | 15.8 | 17.6 |
| Video clustering | 25,200 | 3900 | 3600 | 4.4 | 6.2 |
| User clustering | 24,800 | 3900 | 3600 | 4.4 | 6.2 |

### 5.2. Topic and preference illustration

*Results on YouKu data*: First, we choose the videos with more than 15 keywords and the users who access more than 10 videos from YouKu media data set. The statistics of the data set are shown in the first row of Table 3 where *d*/*u* and *w*/*d* denote the average number of documents that a user has viewed and the average number of words that a document has. We apply PTM to this set and empirically set the number of topics to be 100 and the number of preferences to be 10.

Examples of extracted topics are illustrated in Table 4. Each topic is represented by the top six words (translated from Chinese, removing segmentation noise) most likely to be generated conditioned on the topic. The numbers indicate the probabilities that the words belong to the topics. Thus the topics are the specific representations of the content: news, American comic, a Korea pop group, Hongkong stars, NBA, Chinese treasure programs, cross-talk, science and education programs, which are summarized in the captions.

If a preference is represented by the topics which are most likely to be generated conditioned on the preference, the meaning of the preference can also be illustrated. For example, we find that there is a preference that generates Topic 46, Topic 56, Topic 66, Topic 71 (the topics are shown in Table 4), which means that people with the preference are interested in some funny and fashion videos like cross-talk, pop group, comedy, etc. Another preference generates Topic 31, Topic 36 and Topic 57, which shows that people with the preference pay more attention to the news and society. These two preferences are shown in Fig. 4.

**Table 4**
Topics discovered by Preference-Topic model on YouKu data.

| Topic 31: news | | Topic 56: American comic | | Topic 66: a Korea pop group | |
|---|---|---|---|---|---|
| Terror | 0.068 | comic | 0.046 | Tong Vfang Xien Qi | 0.080 |
| Hot | 0.060 | occident | 0.046 | XIAH (a group member) | 0.045 |
| Violence | 0.055 | Batman | 0.045 | HERO (a group member) | 0.042 |
| News | 0.052 | hero | 0.045 | U-Know (a group member) | 0.042 |
| Event | 0.052 | justice | 0.045 | TVXQ | 0.042 |
| Society | 0.052 | superman | 0.045 | Korea | 0.031 |
| Topic 59: NBA | | Topic 46: cross-talk | | Topic 71: Hongkong stars | |
| NBA | 0.100 | humor | 0.057 | Humor | 0.037 |
| Basketball | 0.100 | cross-talk | 0.044 | Stephen Chow (an actor) | 0.034 |
| Kobe | 0.080 | classic | 0.048 | Andy Lau (an actor) | 0.034 |
| Rockets | 0.057 | Degang Guo (an actor) | 0.037 | Hongkong | 0.034 |
| Yao Ming | 0.050 | Deyun Club | 0.023 | Act the leading role | 0.032 |
| James | 0.040 | wonderful | 0.021 | Star | 0.032 |
| Topic 36: Chinese treasure programs | | Topic 57: Sci. & Edu. programs | | | |
| China | 0.065 | Essay | 0.042 | | |
| Archaeology | 0.062 | TV | 0.023 | | |
| Legend | 0.045 | CCTV | 0.023 | | |
| Discovery | 0.039 | Science education film | 0.022 | | |
| Buried treasure | 0.037 | Newsreel | 0.022 | | |
| Newsreel | 0.032 | Culture | 0.019 | | |

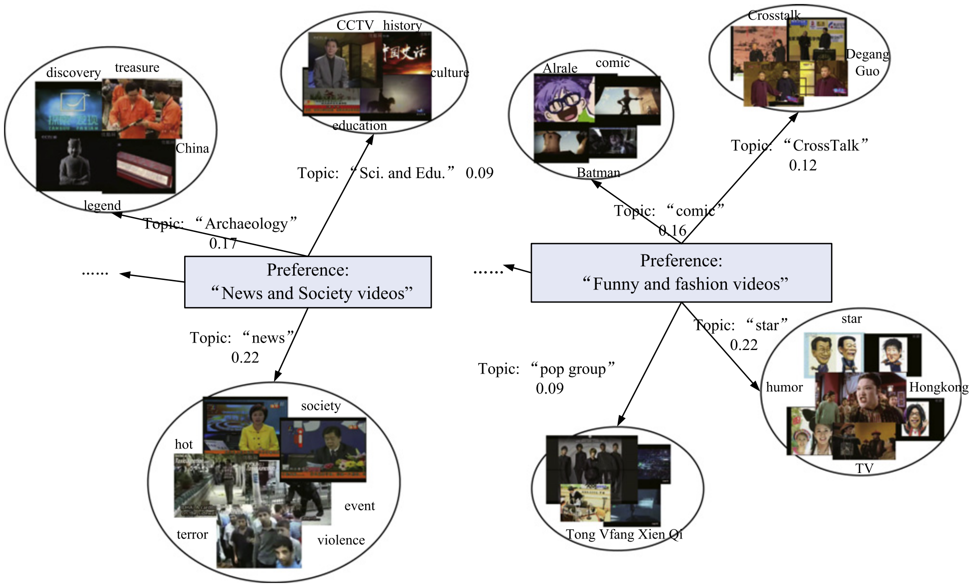

**Fig. 4.** Two preference examples. The number associated with each topic, e.g., 0.22, is the conditional probability that the preference generates the topic.

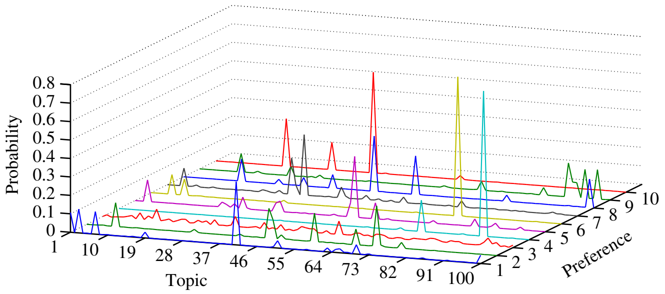

**Fig. 5.** The distribution of preferences over topics on YouKu data.

The lines with arrows represent that preference generates topics, while the number associated with each topic, e.g., 0.22, denotes the conditional probability that the preference generates the topic.

On the other hand, preference distribution can also be used to analyze the correlation between these topics, i.e., subjectively two topics are well-correlated if they are most likely generated by one preference. Thus Fig. 5 shows the correlations between these 100 topics.

*Results on DBLP data*: Table 5 shows examples of preferences and topics from DBLP data. We summarize the preferences in the captions. Thus the preferences obtained from DBLP data set are mainly on the topics related to data mining, machine learning, text mining and search, information retrieval, query analysis and search engine respectively, which are all research topics in these conferences.

Besides topics and preferences, PTM can also obtain users' preference distributions which are indicated by parameter $\gamma$ values in the model. Fig. 6 shows four famous researchers' preference distributions. The figure indicates that Vladimir–Vapnik has the greatest probability on Preference 7 which is related to machine learning; W. Bruce Croft has the greatest probability on Preference 9 which is related to information retrieval while Jiawei Han and Christos Faloutsos have the greatest probability on Preference 5 which is related to data mining. Moreover, the preference distributions of Jiawei Han and Christos Faloutsos are similar, as they are both active researchers in data mining domain.

Based on the preference distributions, we can calculate the interest distance between users. For example, the distances between Jiawei Han and all the other authors who have written more than nine papers are calculated and Table 6 shows the top four researchers with the smallest distances. $n$ is the number of co-authored papers. The results are very telling. Although Mohammed Javeed Zaki did not co-author any papers with Jiawei Han, they are both active on the research topics of graph mining, clustering, etc. Other authors all have cooperated with Jiawei Han.

### 5.3. Video and user clustering performance evaluation

Through inference, the conditional probability $p(z|d)$ that a video belongs to a topic can be estimated. We use this probability to label a video's cluster and evaluate the quality of topics by comparing the obtained clusters with the ground-truth clusters. Intuitively, if the topics are more accurate, the cluster results of videos should be closer to the ground-truth. In the same way, we evaluate the quality of preferences through user clustering.

*Data sets with ground-truth*: We select videos of four categories: basketball, soccer, movie and drama, labeled by the organizers of the website YouKu. Thus we treat them as the ground-truth. The corresponding user access logs are included. This data set is used for video clustering. Moreover, we select four categories of users, who mainly watch basketball, soccer, movie and drama, and include the corresponding video keywords. This data set is used for user clustering. The statistics of the data sets are shown in the second and third rows of Table 3, which demonstrates that they are very sparse.

*Metric*: The clustering results are evaluated by the clustering comparison method in [13]. Given $N$ data objects, the similarity between two clustering results is defined as

$$Pc = \frac{N_{00}+N_{11}}{N(N-1)/2} \tag{14}$$

where $N_{00}$ denotes the count of object pairs that are in different clusters for both clustering methods and $N_{11}$ is the count of pairs that are in the same cluster.

We use this metric to measure the similarity between the clusters inferred from LDA or PTM and the ground-truth clusters. A larger $Pc$ indicates a greater similarity between the clusters

**Table 5**
Preferences discovered from DBLP data set.

| | | |
|---|---|---|
| P5 | | Data mining |
| T53 | 0.128 | Mine system database discovery knowledge text rule large |
| T09 | 0.114 | Data classify summary web-page predict stream analysis mine |
| T49 | 0.088 | Mine pattern data extract answer question frequent period |
| T04 | 0.085 | Data cluster mine high-dimension efficient wide world database |
| T69 | 0.072 | Data mine stream concept-drift issue select correct classify |
| T47 | 0.066 | Mine data gene pattern rule efficient constraint express |
| P7 | | Machine learning |
| T21 | 0.104 | Support vector machine classify regress classifies sparse program |
| T78 | 0.079 | Learn data adapt reinforce structure semi-supervised machine Bayesian |
| T83 | 0.074 | Learn method kernel model active reinforce supervised dynamic |
| T95 | 0.067 | Learn model related cluster control visual motor Markov |
| T70 | 0.059 | Learn cluster model process local Gaussian mixture queries |
| T58 | 0.055 | Learn network approach regular large function approximate linear |
| P8 | | Text mining and search |
| T52 | 0.114 | Document cluster method model select relevant rank graph |
| T30 | 0.100 | Queries search text mine language improve log mobile |
| T10 | 0.084 | Model algorithm structure analysis link factor classify data |
| T65 | 0.081 | Search retrieval system text combine document rank automatic |
| T75 | 0.063 | Search improve mine result model graph corpora index |
| T32 | 0.062 | Search person reduce dimension result evaluate abstract demonstrate |
| T18 | 0.056 | Detect novelty event text comparison track visual evaluate |
| P9 | | Information retrieval and analysis |
| T89 | 0.249 | Retrieval model inform language probabilistic classifier approach method |
| T22 | 0.123 | Retrieval system relevant document feedback evaluate text image |
| T50 | 0.107 | Kernel analysis estimate depend operate linear component principle |
| T40 | 0.087 | Query inform retrieval translate expansion system cross-language phrase |
| P10 | | Query analysis, search engine |
| T26 | 0.075 | Semantic latent index text design analysis product collect |
| T43 | 0.071 | Search engine database annotate application spatial interface image |
| T28 | 0.069 | Query cluster classify search knowledge kernel data limit |
| T86 | 0.068 | Cluster spectral evolutionary vs. kernel temporal neuron curve |
| T70 | 0.062 | Learn cluster model process local gaussian mixture queries |
| T15 | 0.054 | Visual interact converge search automatic user rate person |

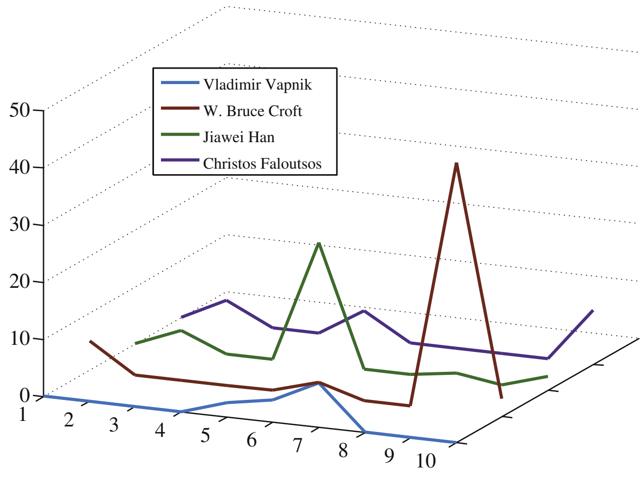

**Fig. 6.** The examples of four researchers' $\gamma$ distributions. The number of preferences is set to be 10 empirically. The details of some preferences are presented in Table 5. It shows that Vladimir Vapnik has the greatest probability on Preference 7 which is related to machine learning; W. Bruce Croft has the greatest probability on Preference 9 which is related to information retrieval while Jiawei Han and Christos Faloutsos have the greatest probability on Preference 5 which is related to data mining.

and the ground-truth. As the cluster labels are inferred based on topics or preferences, a larger *Pc* also indicates better topics or preferences.

*Performance comparison*: First, we cluster videos based on the topics obtained from LDA or PTM respectively. The comparison results are shown in Fig. 7. The evaluations are conducted with the number of topics ranging from 4 to 20. Fig. 7 shows that the clustering results of PTM are better than the clustering results of LDA, therefore it proves that PTM indeed enhances the sparse textual information of videos and obtains better topics than LDA by including user information.

Besides topics, PTM can also discover preferences. As a preference is a higher-level representation over topics, the related topics are likely to be generated conditioned on the same preference. For example, the results in Section 5.2 demonstrate that there is a preference which generates related topics: Topic 31, Topic 36 and Topic 57. In other words, a preference brings the related topics under a common theme at a higher abstract level. Thus we can estimate the conditional probability that a video belongs to a preference as Eq. (15) and use it to infer the video cluster label

$$p(x_i|d) \propto \sum_{j=1}^{T} p(z_j|d) * \mu_{ij} \tag{15}$$

The results shown in Fig. 8(a) demonstrate that the clustering results inferred based on preferences are the best. Thus it proves that preferences connect related topics together and the hierarchial framework of PTM can help get better video clusters on a higher abstract level when the data is sparse.

Besides video clustering, we also cluster users to evaluate the quality of preferences. LDA can also be employed to get preferences on user access logs data [2]. The comparison result shown in Fig. 8(b) demonstrates that PTM is also superior to LDA in clustering users based on the discovered preferences. Thus it indicates that topics can help get better preferences. The reason is that the users who access different but similar-content videos can be identified to have similar interests by the similar topics of the accessed videos in PTM.

In general, as PTM considers both the user logs and the video textual information in a unified model, it enables preferences and topics to benefit each other so as to overcome the problem of sparse data. Therefore PTM is superior to LDA in discovering more informative topics and preferences in terms of clustering-based evaluations.

### 5.4. *Perplexity test*

Perplexity is usually used to estimate the generalization performance of graphical models. In this part, we compare the perplexity of PTM with LDA.

**Table 6**
Symmetric KL divergence of authors.

| Author | $n$ | KL-Dis |
|---|---|---|
| Mohammed Javeed Zaki | 0 | 0.126 |
| Philip S. Yu | 5 | 0.236 |
| Charu C. Aggarwal | 1 | 0.306 |
| Jian Pei | 3 | 0.349 |
| MAXIMUM distance | | 4.28 |

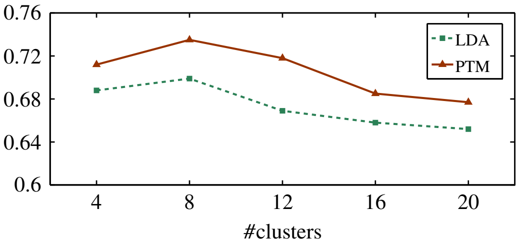

**Fig. 7.** Video cluster similarity comparisons vs. the number of topics ranging from 4 to 20. It demonstrates that the clustering results of PTM are better than the clustering results of LDA.

We conduct two kinds of experiments, in which we treat a portion of videos or users in a corpus as unlabeled test data respectively. Then we compute the perplexity of a held-out test set and compare the results.

PTM or LDA can be used to infer the likelihood of the testing data. A higher likelihood indicates the better estimation on pre-unknown data. As the perplexity is algebraically equivalent to the inverse of the geometric mean per-word likelihood, a lower perplexity score would indicate better generalization performance.

Formally, for a test set of $M$ users, the perplexity is

$$perplexity(D_{test}) = -\frac{\sum_{m=1}^{M}\sum_{n=1}^{N}\log p(\mathbf{w_{mn}})}{\sum_{m=1}^{M}\sum_{n=1}^{N} N_{mn}} \qquad (16)$$

where $\boldsymbol{w}_{\mathbf{mn}}$ is the word sequence of video $n$ which is accessed by user $m$, $N_{mn}$ is the number of words of each video.

Table 7 shows the perplexity values of LDA and PTM at different test proportions. It demonstrates that PTM provides better generalization performance than LDA.

## 6. Related work

*Document topic modeling*: In recent years, many probabilistic topic models have been proposed for various tasks. PLSA [1] and LDA [2] are classic works which are capable of discover latent topics of documents. Later, many researchers have incorporated other information into topic modeling for various applications. For example, the method [3] has been proposed to discover the topic trends over time dimension. Mei [4,14] studied the relationship between topics and geographic locations. Zhou [13] has employed topic models to discover communities with semantic meaning. Some other works have proposed topic models on network data, e.g., using network structure to smooth topic distributions [8,9], predicting citation [15,16], analyzing social influence [17,18]. Multimedia analysis is a very important area which many

**Table 7**
Perplexity comparison.

| Test proportion | | 0.1 | 0.2 | 0.3 |
|---|---|---|---|---|
| Hold out video | LDA | 7.45 | 6.85 | 7.15 |
| | PTM | 6.44 | 6.34 | 6.40 |
| Hold out user | LDA | 19.85 | 22.41 | 21.97 |
| | PTM | 11.45 | 18.39 | 18.25 |

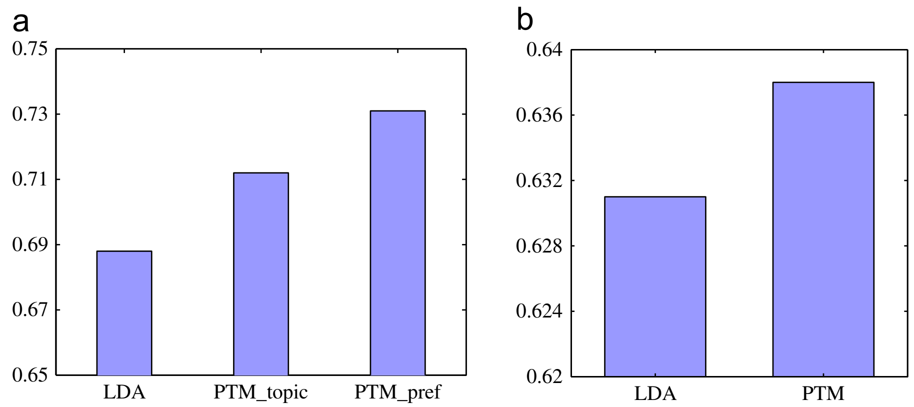

**Fig. 8.** (a) Video cluster similarity comparison. It demonstrates that the clustering results inferred based on preferences are the best; (b) user cluster similarity comparison. It demonstrates that PTM can also get better user clustering result than LDA.

researchers study in [19–22], but none of these models study the problem of mining topics and preference on social media data.

*User interest modeling*: With the development of Web 2.0 technology, mining user interests on the abundant user logs has attracted a considerable amount of interest from both academia and industry. The approaches can be classified as either memory-based or model-based methods. In memory-based approaches [23–25], user access histories are all stored and used as the evidence to predict user interests. So the online computation cost increases significantly when the history data becomes huge. The model-based approaches have been proposed later to overcome the problem. Some generative graphical models [26–30] have been proposed to utilize the pure user access logs instead of textual information. Therefore they suffer from two fundamental problems: sparsity and first-rater problems [31]. It means if a new user comes and has little access logs, his/her interests are difficult to discover at first; if some web documents have never been accessed before, it would also be difficult to be found later. Some researchers have utilized the textual information to discover the user interests, e.g., the work [32]. But it suffers from the over-fitting problem. Author-Topic model (ATM) [5,6] has introduced the factor of authors to infer author interests on bibliographic data, which has been extended by the work [7]. Compared with our model, ATM cannot be used to discover preferences on other kinds of user-document interaction data, e.g., web video data in our experiments, because ATM is based on the assumption that topics are generated by authors, which is untenable in other kinds of user–document interactions. For example, for most real-life video sharing applications, the tags and other textual information are in fact generated by users who upload them, instead of those who browse them. Thus ATM cannot be used to discover meaningful topics and preferences on this data since its assumption is violated.

## 7. Conclusions

In this paper, we proposed a unified probabilistic graphical model, named Preference-Topic model, which combines logs and textual information to mine user preferences and document topics simultaneously on social media data. By considering preferences and topics in a unified framework, PTM enables them to benefit each other and utilizes preferences as higher-level constraints over topics to enhance the insufficient textual descriptive information. The experimental results on web video data demonstrated that PTM is superior to LDA in discovering more informative topics and preferences in terms of clustering-based evaluation. The model's good generalization performance in terms of perplexity has also been demonstrated in the experiments.

## Acknowledgments

The work was supported by the National Natural Science Foundation of China under Grants 61103065, 61003097 and 60933013.

## References

[1] T. Hofmann, Probabilistic latent semantic indexing, in: Proceedings of the 22nd Annual International SIGIR Conference, 2003, pp. 259–266.
[2] D. Blei, A. Ng, M. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[3] X. Wang, A. McCallum, Topics over time: a non-Markov continuous-time model of topical trends, in: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006, pp. 424–433.
[4] Q. Mei, C. Liu, H. Su, C. Zhai, A probabilistic approach to spatio-temporal theme pattern mining on weblogs, in: Proceedings of the 15th international Conference on World Wide Web, 2006, pp. 533–542.
[5] M. Steyvers, P. Smyth, M. Rosen-Zvi, T. Griffiths, Probabilistic author-topic models for information discovery, in: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004, pp. 306–315.
[6] M. Rosen-Zvi, T. Griffiths, M. Steyvers, P. Smyth, The author-topic model for authors and documents, in: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, 2004, pp. 487–494.
[7] A. McCallum, A. Corrada-Emmanuel, X. Wang, The author-recipient-topic model for topic and role discovery in social networks, with application to enron and academic email, in: Proceedings of the Workshop on Link Analysis, Counterterrorism, and Security (in Conjunction with SIAM International Conference on Data Mining), 2005, pp. 33–44.
[8] Q. Mei, D. Cai, D. Zhang, C. Zhai, Topic modeling with network regularization, in: Proceedings of the 17th International Conference on World Wide Web, 2008, pp. 101–110.
[9] Y. Sun, J. Han, J. Gao, Y. Yu, iTopicModel: information network-integrated topic modeling, in: Proceedings of the 2009 IEEE International Conference on Data Mining, 2009.
[10] L. Liu, J. Tang, J. Han, M. Jiang, S. Yang, Mining topic-level influence in heterogeneous networks, in: Proceedings of the 2010 ACM International Conference on Information and Knowledge Management, 2010, pp. 199–208.
[11] T.L. Griffiths, M. Steyvers, Finding scientific topics, Proc. Natl. Acad. Sci. 101 (Suppl. 1) (2004) 5228–5235.
[12] M. Jordan, Z. Ghahramani, T. Jaakkola, L. Saul, Introduction to variational methods for graphical models, Mach. Learn. 37 (1999) 183–233.
[13] D. Zhou, E. Manavoglu, J. Li, C.L. Giles, H. Zha, Probabilistic models for discovering e-communities, in: Proceedings of the 15th international Conference on World Wide Web, 2006, pp. 173–182.
[14] Q. Mei, C. Zhai, A mixture model for contextual text mining, in: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006, pp. 649–655.
[15] J. Tang, J. Zhang, J.X. Yu, Z. Yang, K. Cai, R. Ma, L. Zhang, Z. Su, Topic distributions over links on web, in: Proceedings of the 2009 IEEE International Conference on Data Mining, 2009.
[16] R.M. Nallapati, A. Ahmed, E. Xing, W.W. Cohen, Joint latent topic models for text and citations, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008, pp. 542–550.
[17] J. Tang, J. Sun, C. Wang, Z. Yang, Social influence analysis in large-scale networks, in: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2009.
[18] L. Dietz, S. Bickel, T. Scheffer, Unsupervised prediction of citation influences, in: Proceedings of the 24th International Conference on Machine Learning, 2007, pp. 233–240.
[19] R. Hong, M. Wang, M. Xu, S. Yan, T.-S. Chua, Dynamic caption: video accessibility enhancement for hearing impairment, in: ACM Multimedia, 2010.
[20] R. Hong, M. Wang, X.-T. Yuan, M. Xu, J. Jiang, S. Yan, T.-S. Chua, Video accessibility enhancement for hearing impaired users, ACM Trans. Multimedia Comput. Commun. Appl. 7S (1) (2011).
[21] M. Wang, X. Hua, R. Hong, J. Tang, G. Qi, Y. Song, Unified video annotation via multi-graph learning, IEEE Trans. Circuit Syst. Video Technol. 19 (5) (2009) 733–746.
[22] M. Wang, X.-S. Hua, J. Tang, R. Hong, Beyond distance measurement: constructing neighborhood similarity for video annotation, IEEE Trans. Multimedia 11 (3) (2009) 465–476.
[23] J. Wang, A.P. de Vries, M.J.T. Reinders, Unifying user-based and item-based collaborative filtering approaches by similarity fusion, in: Proceedings of the ACM Conference on Research and Development in Information Retrieval, 2006, pp. 501–508.
[24] J.L. Herlocker, J.A. Konstan, A. Borchers, J. Riedl, An algorithmic framework for performing collaborative filtering, in: Proceedings of the ACM Conference on Research and Development in Information Retrieval, 1999, pp. 230–237.
[25] M. Deshpande, G. Karypis, Item-based top-n recommendation algorithms, ACM Trans. Inf. Syst. 22 (1) (2004) 143–177.
[26] J.S. Breese, D. Heckerman, C. Kadie, Empirical analysis of predictive algorithms for collaborative filtering, in: Proceedings of the 14th Conference, Uncertain in Artificial Intelligence, 1998, pp. 43–52.
[27] T. Hofmann, J. Puzicha, Latent class models for collaborative filtering, in: Proceedings of the International Joint Conference on Artificial Intelligence, 1999, pp. 688–693.
[28] L. Si, R. Jin, A flexible mixture model for collaborative filtering, in: Proceedings of the 20th International Conference on Machine Learning, 2003, pp. 704–711.
[29] R. Jin, L. Si, C. Zhai, Preference-based graphical models for collaborative filtering, in: Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence, 2003.
[30] R. Jin, L. Si, C. Zhai, A study of mixture models for collaborative filtering, J. Inf. Retr. 9 (3) (2006) 357–382.
[31] P. Melville, R.J. Mooney, R. Nagarajan, Content-boosted collaborative filtering for improved recommendations, in: Proceedings of the 18th National Conference on Artificial Intelligence, 2002, pp. 187–192.
[32] A. Popescul, L.H. Ungar, D.M. Pennock, S. Lawrence, Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments, in: Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence, 2001.

**Lu Liu** received her BE degree and PhD degree in the Department of Computer Science and Technology of Tsinghua University in 2005 and 2010, respectively. She is now an Assistant Professor at the Capital Medical University, Beijing, China. Her research interests include multimedia analysis, information retrieval, social network mining, generative graphical model, etc.

**Feida Zhu** is currently an Assistant Professor at the School of Information Systems in Singapore Management University. His research interests are data mining, web mining, algorithms and complexity analysis for data mining and database problems. He got his PhD in Computer Science from University of Illinois at Urbana-Champaign under the supervision of Jiawei Han in 2009. During his PhD study, he has won two Best Student Paper Awards from ICDE (International Conference on Data Engineering Conference) 2007 and PAKDD (The Pacific-Asia Conference on Knowledge Discovery and Data Mining) 2007, respectively.

**Lei Zhang** is a Lead Researcher in the Web Search & Mining Group at Microsoft Research Asia in Beijing, and an Adjunct Professor of Tianjin University. He got his PhD in Computer Science from Tsinghua University in 2001 and then joined Microsoft Research Asia and worked with the Media Computing Group on key projects such as image classification, red-eye detection, face detection and annotation. He is an IEEE and ACM member, and has served as a Program Co-chair of MMM 2010, and served on international conference program committees, including ACM Multimedia, WWW, SIGIR, WSDM, ICME, MMM, PCM, etc. He is the author or co-author of more than 80 published papers in fields such as content-based image retrieval, computer vision, Web search and information retrieval. He also holds 10 U.S. patents for his innovation in face-detection, red-eye reduction and image retrieval technologies.

**Shiqiang Yang** graduated from the Department of Computer Science and Technology, Tsinghua University in 1977 and received the ME degree in 1983. He is now a Professor at Tsinghua University. His research interests include multimedia technology and systems, video compression and streaming, content-based retrieval and semantics for multimedia information, embedded multimedia systems. He has published more than 100 papers in the international conference and journals. Yang is currently the President of the Multimedia Committee of the China Computer Federation. He is a senior member of IEEE.