Inferring user consumption preferences from social media

Yang LI
Harbin Institute of Technology

Jing JIANG
Singapore Management University, jingjiang@smu.edu.sg

Ting LIU
Harbin Institute of Technology

PAPER
Inferring User Consumption Preferences from Social Media

Yang LI[†a)], *Member*, **Jing JIANG**[††b)], *and* **Ting LIU**[†c)], *Nonmembers*

SUMMARY   Social Media has already become a new arena of our lives and involved different aspects of our social presence. Users' personal information and activities on social media presumably reveal their personal interests, which offer great opportunities for many e-commerce applications. In this paper, we propose a principled latent variable model to infer user consumption preferences at the category level (e.g. inferring what categories of products a user would like to buy). Our model naturally links users' published content and following relations on microblogs with their consumption behaviors on e-commerce websites. Experimental results show our model outperforms the state-of-the-art methods significantly in inferring a new user's consumption preference. Our model can also learn meaningful consumption-specific topics automatically.
key words: *social media, e-commerce websites, consumption preferences, topic model*

1. Introduction

Recent years, social media services have become an indispensable part of our daily lives. People express their feelings, needs and desires through online activities like chatting with friends and posting short status updates [1]. Users' personal information and activities on social media presumably reveal their consumption interests. This offers great opportunities for many e-commerce companies to adopt social media as a marketing place. For example, if a user cares much about the price of gasoline or car services, he is likely to own a car and therefore may often buy products from the category of *Car Accessory*. If a user frequently talks about baby and follows famous pediatricians, we can infer that she may like to buy products from the category of *Baby*. It is important to learn user consumption preferences from their social media activities so that better tailored products or services can be recommended [2], [3]. Intuitively, using the consumption preferences, we can recommend products from categories that a user is more interested in, to provide a better user experience and to help increase the probability of clicks.

However, it is still challenging to leverage users' pub-

lished content and following relations across platforms for improving product recommendations. The reasons are two-fold. First, it is hard to link users' identities in social networking sites to their identities on e-commerce sites [4], [5]; Second, not all published content or following relations are reflective of purchase intentions.

Fortunately, e-commerce and social networking have been closely intertwined in recent years. Many e-commerce websites support the mechanism of social login, which allows new users to sign in with their existing login information from social networking services such as Facebook, Twitter and Google+. Users are increasingly encouraged to connect to online social media from e-commerce sites and share their consumption activities with friends. For example, users may share a comment like "I have just bought . . . on eBay. The link is . . ." in their microblog posts. These linking traces provide great potential for e-commerce sites to utilize users' activities on other social media platforms to promote the right kinds of products to the right people.

In this paper, we study how we can leverage users' status updates and following relations on microblogs to recommend the right categories of products to them. We take a supervised approach and propose to use the linkage mined between microblogging sites and e-commerce websites, as illustrated in Fig. 1. Specifically, we propose a principled latent variable model that naturally links users' published content and following relations on microblogs with their consumption behaviors on e-commerce websites. Our model is based on standard LDA. Differently, we assume that each user has a consumption topic distribution and a background topic distribution. We associate every consumption topic with a meta-category label and jointly model the content words as well as the followees of a user.

We build our dataset from the largest Chinese microblogging service Sina Weibo* and a large Chinese e-commerce website Jingdong**, containing a total of 15,186 linked users. Through experimental evaluation, we show the feasibility and effectiveness of our method.

The main contributions of this paper are summarized as follows:

- We propose a consumption preference topic model that naturally links users' published content and following relations on microblogs with their consumption behaviors on e-commerce websites. Our model can be gen-

*http://weibo.com
**http://www.jd.com

eralized to learn a new user's consumption preference.

- We show experimentally that our model outperforms the baselines in inferring users' consumption preferences. Furthermore, our model can learn meaningful words and followee accounts associated with consumption categories.

2. Background

Given a set of users together with their published content and following relations on a microblogging site, our goal is to infer each user's consumption preferences. Specifically, we focus on consumption preferences at the category level, which are represented as an ordered list of a pre-defined set of product categories. We use 12 meta-categories from the e-commerce website Jingdong: *Mobile Phone*, *Baby*, *Beauty*, *Clothes*, *Home&Funiture*, *Footwear&Bags*, *Car Accessory*, *Food&Drinks*, *Pets*, *Sports&Outdoors*, *Household appliances* and *Computer&Digital*.

Semantic Matching. A simple way to solve the problem is to rank the categories relevant to a user based on semantic matching. However, this solution is totally unsupervised and suffers from several shortcomings: (1) Extracting semantic features for a category is inherently not easy. (2) There is likely a discrepancy (word mismatch) between the vocabularies used in user-generated content and category descriptions. For example, a natural way to represent a category is to use the category and product names. However, users rarely express their needs by mentioning a category name or product name explicitly [6], [7]. (3) The naive solution cannot provide an interpretation of users' consumption interests.

Fortunately, we can utilize the linkage mined between microblogging sites and e-commerce websites to solve this problem in a supervised approach. For a user whose purchase history is known, we concatenate all her posts (or followees) into a pseudo document and treat the categories of her purchased items as labels to the document. Hence,

the task becomes learning word-label correspondences and inferring the labels for users based on their content. An immediately available solution to this problem setup is Labeled LDA (L-LDA) [8]. L-LDA is a topic model that constrains LDA by defining a one-to-one correspondence between LDA's latent topics and user labels (or categories in our case). However, for our task, we not only have the category information for a set of "training" users but also the distributions of the categories for these users. Such distributional information is useful and standard L-LDA needs to be modified to allow incorporation of such information.

3. Method

In this section, we first give a general introduction of how to directly apply L-LDA for our task. Then we present the major limitations of L-LDA model and propose our model.

3.1 Labeled LDA Model

A basis of our problem setting is that we assume there is a set of U users and K categories. For each user u, d_u is a pseudo document comprising a list of N_u words. To represent each user's purchase history, we use a binary vector $\Lambda^{(u)} = (l_1, \ldots, l_K)$, where l_k is 1 when user u has purchased some product in category k and 0 otherwise. Next, we define the set of user u's category labels to be $\lambda^{(u)} = \{k | \Lambda_k^{(u)} = 1\}$.

L-LDA assumes that there are K topics corresponding to K categories. The generative process of L-LDA is: First, draw a multinomial topic distribution ϕ_k over the vocabulary for each topic k from a Dirichlet prior β, which is the same as in traditional LDA model. The LDA model then draws a multinomial mixture distribution θ_u over all K topics, for each document d_u, from a Dirichlet prior α. Differently, L-LDA restricts θ_u to be defined only over the topics that correspond to its labels $\lambda^{(u)}$ (categories user u has purchased products from). Since the topics are drawn from this distribution, this restriction ensures all the topic assignments are limited to the document's labels. Without going into the de-
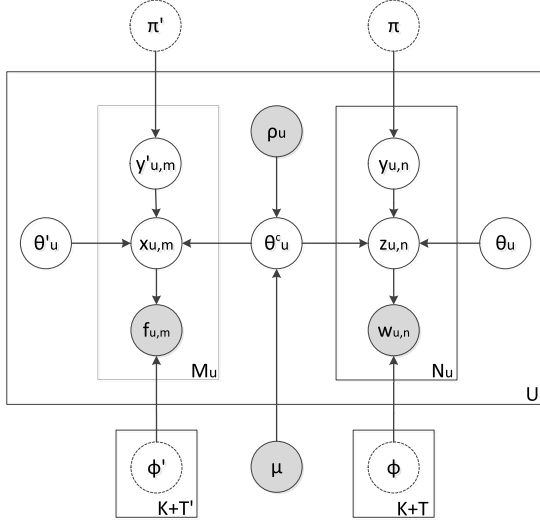
Fig. 2 Plate notation for CPTM. The dashed variables will be collapsed out during Gibbs sampling. Hyperparameters are omitted for clarity.

tails, which can be found in Ramage et al. [8], once the topic distributions are learned from the training set, we can infer topic labels for words from any "new pseudo documents" via Gibbs sampling.

3.2 Our Model

Although L-LDA can be directly used for our task, it has three major limitations.

First, the basic assumption of L-LDA is that the topics of a document are the same as the consumption categories. However, it is not true for our task. In microblogging sites such as Weibo, people talk about not only personal interests that lead to online purchases but also other topics of interest such as opinions on major events, celebrities, etc. Hence in our model, we assume there are generally two kinds of topics, namely **consumption topics** and **background topics**.

The second limitation in L-LDA is, the topic distribution is generated from a symmetric Dirichlet distribution. When directly applied to our task, in the case when a user has bought many products from one category and only one product from another category, the two categories would still be treated equally. This certainly does not make sense. Therefore, in our model, we assume that the **consumption topic distribution** is generated from a Dirichlet distribution parameterized using the user's purchase history over categories, which is *asymmetric*.

Last but not least, the followees of a user can reveal her interests and have correlations with the user's consumption preferences. For example, users who have great interests in *Beauty* may follow those celebrity accounts leading the fashion and beauty trends. In light of this, we propose to incorporate user **followee information** into our model to better profile user interests.

We refer to our proposed model Consumption Preference Topic Model, or CPTM for short (Fig. 2). Our model makes the following assumptions. There exist $K + T$ topics that explain all the posts published by the users, where K is the number of consumption topics (categories) and T is the number of background topics. Each topic has a word distribution φ_k with a uniform Dirichlet prior parameterized by β. We assume that each user has a consumption topic distribution θ_u^c which is related to consumption preferences and a background topic distribution θ_u. For the n-th word $w_{u,n}$ in user u's pseudo document d_u, it can either belong to a consumption topic or a background topic. We use a variable y which is drawn from a Bernoulli distribution to decide the topic type of the word. We also assume that θ_u has a uniform Dirichlet prior parameterized by α, while θ_u^c has a Dirichlet prior parameterized by user specific vector ρ_u.

Now we introduce how to generate the vector ρ_u. Recall that there is a consumption label set $\lambda^{(u)}$ for each user u. H_u is the number of labels $H_u = |\lambda^{(u)}|$. Let $L^{(u)} \in \mathbb{R}^{H_u \times K}$ be a document specific label projection matrix. For each row $i \in \{1, 2, \ldots, H_u\}$ and column $k \in \{1, 2, \ldots, K\}$. We define each entry $L_{i,k}^{(u)}$ as follows:

$$L_{i,k}^{(u)} = \begin{cases} Score(u,k), & \text{if } i\text{-th label corresponds to topic } k \\ 0, & \text{otherwise} \end{cases}$$

where k is a consumption topic, and its corresponding consumption category is c_k. $Score(u,k)$ is the number of products purchased by the user in category c_k normalized by the number of products from all categories in user u's purchase history. We have

$$Score(u,k) = \frac{purc(u, c_k)}{\sum_{i \in K} purc(u, c_i)}, \qquad (1)$$

Then ρ_u is defined in the following formula:

$$\rho_u = L^{(u)} \mu \qquad (2)$$

where $\mu = (\mu_1, \ldots, \mu_K)$ is a K−dimensional parameter vector predefined. Note that in our problem, all the μ_k are the same.

Note that besides words, our model also incorporate users' following relations. Similarly, for every user, we also treat her followees as a pseudo document d_u'. The m-th followee of user u, referred to as $f_{u,m}$, has a topic label $x_{u,m}$. The topic label is generated either from the shared consumption topic distribution θ_u^c or a separate background topic distribution θ_u'. Similar to $z_{u,n}$, we use a switch $y_{u,m}'$ to decide $x_{u,m}$'s topic type. The generative process of CPTM is shown in Algorithm 1.

3.3 Learning and Inference

We use collapsed Gibbs sampling to learn model variables for our model. There are two sets of hidden variables related to two parts of our model, one for modeling user-generated text and the other for modeling user followees. For the first part, for each word $w_{u,n}$, we jointly sample its topic $z_{u,n}$ and the switch variable $y_{u,n}$ (deciding between consumption

Algorithm 1 Generative Process for CPTM.

1: Draw $\pi, \pi' \sim Beta(\gamma)$
2: **for all** topic $k = 1, \ldots, K$ **do**
3:    Draw word probability $\varphi_k \sim \text{Dir}(\beta)$
4:    Draw followee probability $\varphi'_k \sim \text{Dir}(\beta')$
5: **end for**
6: **for all** user $u = 1, \ldots, U$ **do**
7:    Draw consumption topic probability $\theta^c_u \sim \text{Dir}(\rho_u)$
8:    Draw background topic probability $\theta_u \sim \text{Dir}(\alpha)$
9:    Draw background topic probability $\theta'_u \sim \text{Dir}(\alpha')$
10:    **for all** word $w_{u,n} = 1, \ldots, N_u$ of user u **do**
11:       Draw $y_{u,n} \sim \text{Bernoulli}(\pi)$
12:       **if** $y_{u,n} = 1$ **then**
13:          Draw topic $z_{u,n} \sim \text{Multi}(\theta^c_u)$
14:       **else**
15:          Draw topic $z_{u,n} \sim \text{Multi}(\theta_u)$
16:       **end if**
17:       Draw word $w_{u,n} \sim \text{Multi}(\varphi_{z_{u,n}})$
18:    **end for**
19:    **for all** followee $f_{u,m} = 1, \ldots, M_u$ of user u **do**
20:       Draw $y'_{u,m} \sim \text{Bernoulli}(\pi')$
21:       **if** $y'_{u,m} = 1$ **then**
22:          Draw topic $x_{u,m} \sim \text{Multi}(\theta^c_u)$
23:       **else**
24:          Draw topic $x_{u,m} \sim \text{Multi}(\theta'_u)$
25:       **end if**
26:       Draw followee $f_{u,m} \sim \text{Multi}(\varphi_{x_{u,m}})$
27:    **end for**
28: **end for**

topic or background topic).

The probability of assigning a word $w_{u,n}$ to a background topic is defined as:

$$P(y_{u,n} = 0, z_{u,n} = k | Y_{\neg(u,n)}, Z_{\neg(u,n)}, W) \quad (3)$$

$$\propto \frac{N_{u,0,k}^{\neg(u,n)} + \alpha}{N_{u,0,*}^{\neg(u,n)} + T\alpha} \cdot \frac{M_{0,k,w_{u,n}}^{\neg(u,n)} + \beta}{M_{0,k,*}^{\neg(u,n)} + V\beta} \cdot \frac{C_0^{\neg(u,n)} + \gamma}{C_*^{\neg(u,n)} + 2\gamma}$$

where $N_{u,0,k}^{\neg(u,n)}$ is the number of user u's words that are assigned to background topic k, $M_{0,k,w_{u,n}}^{\neg(u,n)}$ is the number of word $w_{u,n}$ assigned to background topic k, $C_0^{\neg(u,n)}$ is the total number of background words. All counters are calculated with the current word $w_{u,n}$ excluded. A missing subscript (e.g. $N_{u,0,*}^{\neg(u,n)}$) indicates a summation over that dimension.

The probability of assigning a word $w_{u,n}$ to a consumption topic is defined as:

$$P(y_{u,n} = 1, z_{u,n} = k | Y_{\neg(u,n)}, Z_{\neg(u,n)}, W) \quad (4)$$

$$\propto \frac{N_{u,1,k}^{\neg(u,n)} + N'_{u,1,k} + \rho_{u,i}}{N_{u,1,*}^{\neg(u,n)} + N'_{u,1,*} + \mu} \cdot \frac{M_{1,k,w_{u,n}}^{\neg(u,n)} + \beta}{M_{1,k,*}^{\neg(u,n)} + V\beta} \cdot \frac{C_1^{\neg(u,n)} + \gamma}{C_*^{\neg(u,n)} + 2\gamma}$$

where the i-th label from user u corresponds to topic k, $N_{u,1,k}^{\neg(u,n)}$ is the number of user u's words that are assigned to consumption topic k, $N'_{u,1,k}$ is the number of user u's followees that are assigned to consumption topic k, $M_{1,k,w_{u,n}}^{\neg(u,n)}$ is the number of word $w_{u,n}$ assigned to consumption topic k, $C_1^{\neg(u,n)}$ is the total number of consumption words. All counters are calculated with the current word $w_{u,n}$ excluded.

For the second part, we jointly sample the topic $x_{u,m}$ and the switch variable $y'_{u,m}$ for each followee in a similar way.

Once the topic distributions are learned from the training set, we can infer topic labels for words from any "new pseudo documents" using Gibbs sampling. Similar inference approach is used in L-LDA [8]. After that, we can learn a consumption topic distribution θ^c_u for a new user, based on which we can recommend categories of products to the user, as discussed in the next section.

4. Experiments

In this section, we first introduce our dataset. We then design experiments to evaluate our proposed method. In the quantitative evaluation, we compare our model with a few baselines in inferring user consumption preferences, while in the qualitative evaluation, we show our model can effectively learn consumption words and followees.

4.1 Data Set

Our task requires data from both an e-commerce website and a social networking site.

E-commerce Data. We used a large Jingdong dataset shared by [9], which is constructed by crawling 138.9 million reviews of 0.2 million products from 12 million users during the whole year of 2013. Each review document would correspond to a unique transaction record. Also, each transaction record consists of a user ID, a product ID and the purchase timestamp. We grouped transaction records by user IDs and obtained a list of purchased products for each user.

Microblogging Data. We collected 5 million active users and their 1.7 billion microblog posts from the largest Chinese microblogging site Sina Weibo during the period from January 2013 to June 2013.

It is observed that many e-commence companies encourage users to share their purchase record on their microblogs via a system-generated short URL, which links to the corresponding product entry on Jingdong. We randomly collected active users on Sina Weibo who have explicitly shared their reviews in their posts by searching using some keywords (e.g. "京东商城网购评价 *(Reviews for shopping in Jingdong Mall)*", "京东让红包飞 *(Money flies, Jingdong Mall)*") before June 2013. By following the URL link, we can obtain the Jingdong account of the Weibo user. In this way, we were able to successfully match 23,917 linked users. Then we obtained the information of these users including post content and followee relations from the microblog data. We removed the users who have fewer than 5 purchase records, 10 posts or 20 followees.

Finally, we totally get 15,186 users as our linked users. For each user, we have a set of posts, a set of followees and a list of items purchased from Jingdong (with item name and category). The statistics of the dataset are summarized in Table 1.

Table 1   Statistics of our datasets.

# linked users	15,186
# posts	1,755,524
# following relations	1,164,411
# purchase records	888,039

4.2   Experimental Settings

4.2.1   Baseline Methods

We consider the following representative baseline methods for comparison.

Semantic Matching (S-Match): This is the naive solution we mentioned in Sect. 2. We directly recommend categories to a user based on the cosine similarity between term vectors representing the user and a category with TF-IDF weighting. Specially, we use all the product and sub-category names to represent a category. We use user's published content and followees to represent a user.

Popularity Ranking (PR): For each user, we recommend categories to her based on the popularity of categories. We use the number of users who have ever purchased from the category in the whole data set to represent its popularity.

Labeled latent Dirichlet Allocation (L-LDA): As stated in previous section, L-LDA is a natural baseline of our model. Note that in L-LDA, each topic is associated with a category, so for each user, we rank the categories based on their topic distribution directly.

Support Vector Machines (SVM): We use SVM^{light} to build an SVM classification model [10] for each Jingdong category c. For training, positive examples are users who buy at least one item in c. During testing, for each unknown user SVM returns a confidence score that we use for ranking. SVM parameters are chosen by grid search on a subset of the training sets. We use user's post content and followee accounts as bag-of-words features.

Consumption Preference Topic Model (CPTM): Our proposed model. During the training stage, we obtain the word (followee) distributions per topic. During testing, for a given test user, we assign topic labels to each word (followee) in her microblog posts and then apply the learned consumption topic distributions to infer her consumption preferences.

4.2.2   Evaluation Metrics

For each user u in our data set, we have obtained her purchase records. We treat the ground truth ranking of categories by assigning to each category c_i the ranking score $Score(u, i)$ according to Eq. (1), and establishing the rank as follows:

$$c_i > c_j \Leftrightarrow Score(u, i) > Score(u, j).$$

For example, if a user buys 3 products in *Food&Drinks*, 2 products in *Home&Funiture*, the gold-standard ranking for the user will be: *Food&Drinks > Home&Funiture*.

The ideal prediction algorithm should provide in output for each user a category ranking equivalent to the ground truth. Therefore, instead of looking at binary predictions and measuring prediction errors, we care more about the quality of the top-ranked categories. In this case, we consider two ranking based measure *NDCG@k* [11] and *P@k* which is similar to Zhang et al. [12].

NDCG@k: For each user we define DCG at position k as:

$$DCG@k = \sum_{i=1}^{k} \frac{weight(i)}{log(i + 1)}$$

where $weight(i)$ is the relevance weight of the category ranked in position i by the algorithm. We set the relevance weight as the proportion of the products a user purchase on that category. We also define IDCG at position k as the DCG of the ground truth at k. Finally, NDCG at position k is thus defined as:

$$NDCG@k = \frac{DCG@k}{IDCG@k}$$

P@k: For each user we define precision at position k as:

$$P@k = \frac{\sum_{i=1}^{k} F(c_i)}{k}$$

where $F(c_i)$ equals to 1 if c_i is the *top*-1 category in user's real purchase ranking list (ground truth), which means c_i is the favorite category of the user. Note that we only consider the *top*-1 category in the real purchase ranking list as "relevant results". The reasons are twofold. First, it is more practical in real scenarios. Actually, we found on average users tend to buy nearly 50% products from their favorite category (*top*-1 category). Second, in this problem, the aim of our work is to infer users' consumption preference which should be focused. We have also tested treating *top*-2 and *top*-3 categories as "relevant", and found our method significantly outperforms others in most cases.

4.2.3   Experimental Setup

We evaluate our models and the baseline methods using 5-fold cross validation. We first randomly divide the data into 5 subsets. In each run we use 4 subsets as training data and half of the 5th subset for validation to tune the hyper-parameters. Then we use the other half of the 5th subset for testing. Finally, we compute *NDCG@k* and *P@k* for each fold by averaging the measures over all testing users.

Following previous work [13]–[15], the Dirichlet hyper-parameters α and β are set to values $50/T$ and 0.01 respectively, where T denotes the number of background topics. Similarly, we set $\mu = T\alpha = 50$ (Eq. (3), (4)).

For our model and L-LDA, we set the number of consumption topics K to 12, which equals the total number of categories. Note that in our model, besides consumption topics, we also have background topics. We test with different number of background topics and find $T = T' = 8$ to

Table 2 Results comparison on *NDCG@k*.

Method	NDCG@1	NDCG@2	NDCG@3	NDCG@4	NDCG@5
S-Match	0.3335	0.3695	0.4021	0.4295	0.4513
PR	0.3784†	0.4219†	0.4494†	0.4713†	0.4957†
L-LDA	0.3705	0.4125	0.4509†	0.4829†	0.5077†
SVM	0.3980†	0.4375†	0.4773†	0.5142†	0.5314†
CPTM-C	0.4380†	0.4732†	0.5032†	0.5288†	0.5497
CPTM	**0.4489†**	**0.4801†**	**0.5092**	**0.5354**	**0.5551**

Table 3 Results comparison on *P@k*.

Method	*P@1*	*P@2*	*P@3*	*P@4*	*P@5*
S-Match	0.1593	0.1447	0.1325	0.1236	0.1161
PR	0.1050	0.1272	0.1182	0.1185	0.1180
L-LDA	0.1563†	0.1540†	0.1460†	0.1379†	0.1305†
SVM	0.1629†	0.1434	0.1377	0.1363	0.1335
CPTM-C	**0.2290†**	0.1856†	0.1620†	0.1467†	0.1370
CPTM	0.2383	**0.1935†**	**0.1669†**	**0.1504†**	**0.1377**

Fig. 3 *NDCG@k* results of different methods when we vary *k* from 1 to 5.

Fig. 4 *P@k* results of different methods when we vary *k* from 1 to 5.

be an optimal setting. For the hyper-parameter of Bernoulli distribution which word topic type y is drawn from, γ is set to be 10. We did not observe significant improvements when γ is set to different values. The detailed results will be discussed in Sect. 4.3.3.

4.3 Quantitative Evaluation

4.3.1 Comparison with Baselines

Table 2 and Table 3 show the results of different methods on the *NDCG@k* and *P@k* metrics. The performance of the naive solution S-Match is very poor, even worse than popularity Baseline on *NDCG@k*, which is due to the discrepancy between the vocabulary used in user content and category content. Utilizing the linking traces and users' published content in a supervised way, both L-LDA and SVM perform better than the popularity baseline. However, the improvement of L-LDA is not substantial. L-LDA assumes all words are used to generate user's consumption topic distribution which directly used to predict purchase behaviors. In fact, users' post contents are not all related to consumption behaviors. CPTM is significantly better than the previous methods, meaning that our model can learn a more effective topic distribution for consumption preferences prediction. †means the result is better than previous method above in the same column at 5% significance level by Wilcoxon signed rank test.

We show the *NDCG@k* and *P@k* curves of different methods in Fig. 3 and Fig. 4. Our model performs the best when *k* varies from 1 to 5. Specially, in Fig. 4, our model achieves a much higher relative improvement comparing with other baselines when *k* is smaller than or equal to 3. In our dataset, we found users tend to buy more than 86% products from their *top*-3 interested categories. This can explain why the differences between methods are smaller when *k* is larger than 3. The observation also demonstrates that users' consumption preferences on categories are focused.
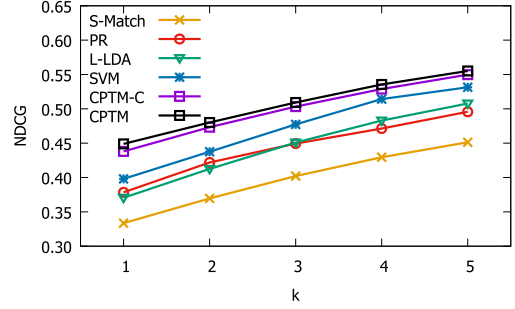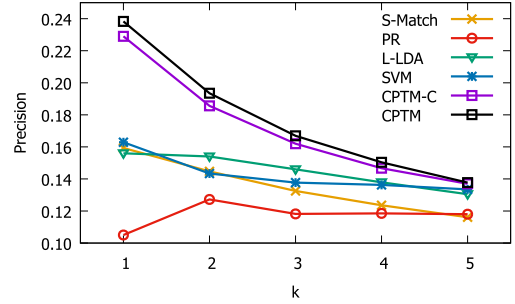
4.3.2 Effectiveness of Following Relations

To verify the effectiveness of following relations, we also compare a degenerate version of our model CPTM-C only incorporating words from users' published content. From Table 2 and Table 3, we find CPTM achieves better results than CPTM-C by incorporating followees additionally. This shows that besides post content, users' followee relations can also indicate their purchase behaviors.

4.3.3 Parameter Sensitive Analysis

We would like to analyze how sensitive the performance of our model is with regard to the parameters.

First, we test the performance of CPTM with number of background topics T ranging from 2 to 20 with a gap of 2. We set $T' = T$. In Fig. 5, we can see CPTM performs best when T is set to 8. The results of *NDCG@k* become flattened when T is larger.

Then we vary the value of hyper-parameter γ while fixing the other parameters. We consider γ for the values: 0.1, 1 and 10. Figure 6 shows the results of CPTM in terms of *NDCG@k*. From Figure 6, we did not observe significant improvements when γ is set to different values, showing that CPTM is not sensitive to γ. We set $\gamma = 10$ in the experiments because it is an optimal setting in most cases.

4.4 Qualitative Evaluation

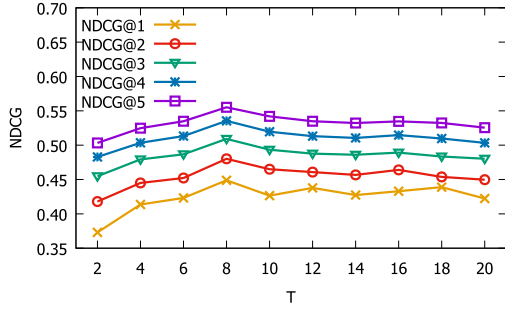We show the top ten words and followee accounts of two

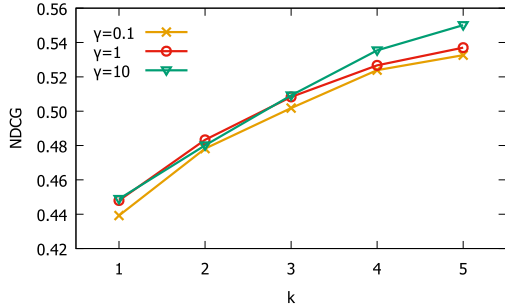**Fig. 5** *NDCG@k* results of CPTM with different number of background topics.

Fig. 6 *NDCG@k* results of CPTM with different γ.

Table 4 Topical words and followee accounts learned from CPTM.

Category		Topical words and followee accounts
Baby	words	宝宝 (baby), 孩子 (child), 妈妈 (mother), 宝贝 (baby), 婴儿 (baby), 儿童 (child), 奶粉 (baby formula), 爸爸 (father), 母婴 (Maternity), 闲置 (unused)
	followee accounts	崔玉涛 (Doctor Cui Yutao), 张思莱医师 (Doctor Zhang Silai), 育儿网官方微博 (Child Care), 好奇官方微博 (Huggies Club), 鲍秀兰 (Doctor Bao Xiulan), 王人平 (Doctor Wang Renping), 京东母婴 (Jingdong Baby), 春雨医生 (Doctor ChunYu), 宝宝树育儿网 (BabyTree)
Beauty	words	时尚 (fashion), 美丽 (beauty), 老公 (husband), 青春 (youth), 韩国 (Korea), 漂亮 (pretty), 浪漫 (romantic), 妈妈 (mother), 衣服 (clothes), 明星 (super star)
	followee accounts	turbosun (Actress Sun Li), 乐蜂网 (Lefeng), 李易峰 (Actor Li Yifeng), 钟汉良 (Actor Zhong Hanliang), 堆糖 (Duitang), in官方微博 (IN-app), 爱图购时尚搭配 (Fashion Collocation), 999款潮流发型 (999 fashion style), veggieg (Actress Wang Fei), 私人挑款师 (Personal stylist)

sampled consumption topics: "*Baby*" and "*Beauty*" in Table 4. Examination shows our model can detect consumption related words and weibo accounts. For example, in "*Beauty*" category, the topic words are all related to "beauty" and "fashion", the followees are either fashion stars or fashion shopping sites such as "Lefeng". In addition, in "*Baby*" category, the words are all related to baby products, the followees include many famous pediatrician accounts such as "崔玉涛 (Doctor Cui Yutao)", "张思莱医师

(Doctor Zhang Silai)" and so on.

Our model differs from vanilla topic models by distinguishing general background topics and consumption topics. Actually from the qualitative evaluation, we found it is necessary to distinguish the two kinds of topics. For example, some general topics we learned are about insights on life and news events, which are talked by almost all the people. We cannot learn any consumption preference from these topics.

Interestingly, we find in the category "*Household appliances*", the learnt top words are something related to family such as "parents", "the elderly" and "children". On the one hand, around 85% of users purchased items from "*Household appliances*" category, making it serve as a background topic and attract many common words. On the other hand, users rarely talk about *Household appliances* on Weibo although they might purchase from it. However, in this case, the followee accounts learned from our model such as "京东家电 *(Jingdong Household appliances)*", "海信电视官方微博 *(Hisense TV)*" can be useful.

5. Related Work

5.1 Commercial Intention Mining

Mining users' commercial intention from social media has attracted much attention theses years [16]–[20]. Wang et al. [21] investigated the use of microblogs data to identify trend-driven commercial intents and trend related products. Hollerit et al. [2] proposed to identify tweet-level commercial intents employing n-grams and part-of-speech tags as features. As an extension of this work, Wang et al. [6] proposed a semi-supervised learning approach to classify intent tweets into six categories. Since most of user needs are implicitly expressed, Ding et al. [7] proposed a domain adaptive convolution neural network to identify the implicit commercial intent tweets. Our work is related to commercial intention mining from social media. However, different from all the existing work identifying commercial intents in a single *tweet*, we focus on studying consumption preferences at *user* level. The tweet level commercial intents are sparse [6], [7] and they do not necessarily link to users' consumption behaviors.

5.2 Product Recommendation from Social Media

Product recommender systems are widely used by e-commerce companies. Recently, much effort has been taken to incorporate various information [22]–[26] to help improve recommendation performance for cold start users, especially from the social media. Zhang et al. [27] and studied interactions between Facebook profiles and purchase behaviors on eBay and showed that users' Facebook profiles can be used to predict online purchase brands. Zhao et al. [3] built a real time recommendation system on microblogs incorporating users' demographic information extracted from their public profiles and product demograph-

ics in e-commerce websites. Using users' demographic attributes in Facebook, Zhang et al. [12] tried to predict user consumption preferences on product categories in eBay to improve "cold start" recommendation. Nevertheless, with privacy concerns, users may not reveal their true demographic attributes, or such attributes may not be made public [5], [28]. Our problem is similar to theirs. However, differently, we focus on leveraging users' published content and following relations, which are public, continuously updated and easier to obtain.

5.3 Topic Models on Microblogs

There has been much interest in topic modeling on microblogs [28]–[31]. Hong and Davison [32] empirically evaluated the performance of Latent Dirichlet Allocation (LDA) [33] on Twitter. They found that by aggregating all tweets published by the same user into a single document the learned topics have higher quality. Zhao et al. [34] proposed Twitter-LDA, which assigns a single topic to an entire tweet. Labeled LDA (L-LDA) [8] is another extension of LDA which assumes that each document has a set of known labels. Ramage et al. [35] proposed to use L-LDA model to characterize user's topics in Twitter.

While designing latent variable models to jointly model content words together with other types of data is not new, to the best of our knowledge, jointly modeling text, social relations and consumption behaviors *across platforms* is new. Inspired by the existing work, we organize all tweets published by the same user into a single document and built our model based on L-LDA, but differently, we introduce two types of topics, namely consumption topics and background topics. Furthermore, we incorporate asymmetric Dirichlet priors to model users' different preferences on consumption categories. Finally, to better profile users, we jointly model user's published content and following relations, which is different from the above discussed models.

6. Conclusions

In this paper, we make the first attempt to infer user consumption preferences from social media in a topic view, which can potentially benefit a multitude of commercial applications. To verify the hypotheses that users' published content and following relations can indicate their online consumption preferences, we design a latent variable model that joint models users' lexical content of posts, following relations and their consumption behaviors on the e-commerce sites. Through experiments, we show that our model outperforms the baseline methods effectively in inferring user consumption preferences. Furthermore, our model is able to discover meaningful words and followee accounts associated with consumption meta-categories.

References

[1] D. Preoţiuc-Pietro, V. Lampos, and N. Aletras, "An analysis of the user occupational class through twitter content," The Association for Computational Linguistics, pp.1754–1764, 2015.

[2] B. Hollerit, M. Kröll, and M. Strohmaier, "Towards linking buyers and sellers: detecting commercial intent on twitter," Proceedings of the 22nd international conference on World Wide Web companion, pp.629–632, International World Wide Web Conferences Steering Committee, 2013.

[3] X.W. Zhao, Y. Guo, Y. He, H. Jiang, Y. Wu, and X. Li, "We know what you want to buy: a demographic-based system for product recommendation on microblogs," Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pp.1935–1944, ACM, 2014.

[4] J. Liu, F. Zhang, X. Song, Y.-I. Song, C.-Y. Lin, and H.-W. Hon, "What's in a name?: an unsupervised approach to link users across communities," Proceedings of the sixth ACM international conference on Web search and data mining, pp.495–504, ACM, 2013.

[5] F. Zhang, N.J. Yuan, D. Lian, and X. Xie, "Mining novelty-seeking trait across heterogeneous domains," Proceedings of the 23rd international conference on World wide web, pp.373–384, ACM, 2014.

[6] J. Wang, G. Cong, X.W. Zhao, and X. Li, "Mining user intents in twitter: A semi-supervised approach to inferring intent categories for tweets," Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015.

[7] X. Ding, T. Liu, J. Duan, and J.Y. Nie, "Mining user consumption intention from social media using domain adaptive convolutional neural network," Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015.

[8] D. Ramage, D. Hall, R. Nallapati, and C.D. Manning, "Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora," Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1, pp.248–256, Association for Computational Linguistics, 2009.

[9] J. Wang, W.X. Zhao, Y. He, and X. Li, "Leveraging product adopter information from online reviews for product recommendation," Ninth International AAAI Conference on Web and Social Media, 2015.

[10] M.A. Hearst, S.T. Dumais, E. Osman, J. Platt, and B. Scholkopf, "Support vector machines," Intelligent Systems and their Applications, IEEE, vol.13, no.4, pp.18–28, 1998.

[11] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of ir techniques," ACM Transactions on Information Systems (TOIS), vol.20, no.4, pp.422–446, 2002.

[12] Y. Zhang and M. Pennacchiotti, "Predicting purchase behaviors from social media," Proceedings of the 22nd international conference on World Wide Web, pp.1521–1532, International World Wide Web Conferences Steering Committee, pp.1521–1532, 2013.

[13] T.L. Griffiths and M. Steyvers, "Finding scientific topics," Proceedings of the National Academy of Sciences, vol.101, no.Suppl. 1, pp.5228–5235, April 2004.

[14] W.X. Zhao, J. Jiang, J. He, Y. Song, P. Achananuparp, E.P. Lim, and X. Li, "Topical keyphrase extraction from twitter," Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, HLT '11, Stroudsburg, PA, USA, pp.379–388, Association for Computational Linguistics, 2011.

[15] Q. Diao, J. Jiang, F. Zhu, and E.P. Lim, "Finding bursty topics from

microblogs," Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1, ACL '12, Stroudsburg, PA, USA, pp.536–544, Association for Computational Linguistics, 2012.

[16] A.B. Goldberg, N. Fillmore, D. Andrzejewski, Z. Xu, B. Gibson, and X. Zhu, "May all your wishes come true: A study of wishes and how to recognize them," Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp.263–271, Association for Computational Linguistics, 2009.

[17] B. Fu and T. Liu, "Weakly-supervised consumption intent detection in microblogs," Journal of Computational Information Systems 6, pp.2423–2431, 2013.

[18] W.Y. Wang, E. Lin, and J. Kominek, "This text has the scent of starbucks: A laplacian structured sparsity model for computational branding analytics," Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013), Seattle, WA, USA, 2013.

[19] H. Amiri and H. Daume III, "Target-dependent churn classification in microblogs," Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015.

[20] C. Yang, S. Pan, J. Mahmud, H. Yang, and P. Srinivasan, "Using personal traits for brand preference prediction," Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, pp.86–96, Association for Computational Linguistics, Sept. 2015.

[21] J. Wang, W.X. Zhao, H. Wei, H. Yan, and X. Li, "Mining new business opportunities: Identifying trend related products by leveraging commercial intents from microblogs," EMNLP, pp.1337–1347, 2013.

[22] S. Aciar, D. Zhang, S. Simoff, and J. Debenham, "Recommender system based on consumer product reviews," Proceedings of the 2006 IEEE/WIC/ACM international Conference on Web intelligence, pp.719–723, IEEE Computer Society, 2006.

[23] M. Giering, "Retail sales prediction and item recommendations using customer demographics at store level," ACM SIGKDD Explorations Newsletter, vol.10, no.2, pp.84–89, 2008.

[24] L. Qiu and I. Benbasat, "A study of demographic embodiments of product recommendation agents in electronic commerce," International Journal of Human-Computer Studies, vol.68, no.10, pp.669–688, 2010.

[25] J. Wang and Y. Zhang, "Opportunity model for e-commerce recommendation: right product; right time," Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, pp.303–312, ACM, 2013.

[26] W.X. Zhao, S. Li, Y. He, L. Wang, J.-R. Wen, and X. Li, "Exploring demographic information in social media for product recommendation," Knowledge and Information Systems, vol.49, no.1, pp.61–89, 2016.

[27] Y. Zhang and M. Pennacchiotti, "Recommending branded products from social media," Proceedings of the 7th ACM conference on Recommender systems, pp.77–84, ACM, 2013.

[28] L. Liao, J. Jiang, Y. Ding, H. Huang, and E.P. LIM, "Lifetime lexical variation in social media," Twenty-Eighth AAAI Conference on Artificial Intelligence, 2014.

[29] J. Eisenstein, B. O'Connor, N.A. Smith, and E.P. Xing, "A latent variable model for geographic lexical variation," Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp.1277–1287, Association for Computational Linguistics, 2010.

[30] Q. Diao and J. Jiang, "A unified model for topics, events and users on twitter," Conference on Empirical Methods in Natural Language Processing, 2013.

[31] M. Qiu, F. Zhu, and J. Jiang, "It is not just what we say, but how we say them: Lda-based behavior-topic model," SIAM.

[32] L. Hong and B.D. Davison, "Empirical study of topic modeling in twitter," Proceedings of the First Workshop on Social Media Analytics, pp.80–88, ACM, 2010.

[33] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent dirichlet allocation," Journal of Machine Learning research, vol.3, pp.993–1022, 2003.

[34] W.X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, "Comparing twitter and traditional media using topic models," Advances in Information Retrieval, pp.338–349, Springer, 2011.

[35] D. Ramage, S.T. Dumais, and D.J. Liebling, "Characterizing microblogs with topic models," ICWSM, vol.10, pp.1–1, 2010.

Yang Li received the Master's degree in July 2012 from the Department of Software Engineering, Harbin Institute of Technology, Harbin, China. Since 2012, she has been a Ph.D. candidate at the Department of Computer Science, Harbin Institute of Technology. Her current research interests include information retrieval, natural language processing, and social media analysis.

Jing Jiang received her Ph.D. degree in 2008 from the Department of Computer Science, University of Illinois at Urbana-Champaign, Illinois, USA. She is an associate professor in the School of Information Systems at the Singapore Management University. Her research interests include natural language processing, information extraction and social media analysis.

Ting Liu received his Ph.D. degree in 1998 from the Department of Computer Science, Harbin Institute of Technology, Harbin, China. He is a Full Professor in the Department of Computer Science, and the Director of the Research Center for Social Computing and Information Retrieval (HIT-SCIR) from Harbin Institute of Technology. His research interests include information retrieval, natural language processing, and social media analysis.