Tu CC, Liu ZY, Sun MS. Tag correspondence model for user tag suggestion. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 30(5): 1063–1072 Sept. 2015. DOI 10.1007/s11390-015-1582-6

Tag Correspondence Model for User Tag Suggestion

Cun-Chao Tu (涂存超), *Student Member, CCF*, Zhi-Yuan Liu* (刘知远), *Senior Member, CCF* and Mao-Song Sun (孙茂松), *Senior Member, CCF*

Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
State Key Laboratory on Intelligent Technology and Systems, Tsinghua University, Beijing 100084, China
National Laboratory for Information Science and Technology, Tsinghua University, Beijing 100084, China
Jiangsu Collaborative Innovation Center for Language Ability, Jiangsu Normal University, Xuzhou 221009, China

E-mail: tucunchao@gmail.com; liuzy@tsinghua.edu.cn; sms@mail.tsinghua.edu.cn

Received November 15, 2014; revised May 12, 2015.

Abstract Some microblog services encourage users to annotate themselves with multiple tags, indicating their attributes and interests. User tags play an important role for personalized recommendation and information retrieval. In order to better understand the semantics of user tags, we propose Tag Correspondence Model (TCM) to identify complex correspondences of tags from the rich context of microblog users. The correspondence of a tag is referred to as a unique element in the context which is semantically correlated with this tag. In TCM, we divide the context of a microblog user into various sources (such as short messages, user profile, and neighbors). With a collection of users with annotated tags, TCM can automatically learn the correspondences of user tags from multiple sources. With the learned correspondences, we are able to interpret implicit semantics of tags. Moreover, for the users who have not annotated any tags, TCM can suggest tags according to users' context information. Extensive experiments on a real-world dataset demonstrate that our method can efficiently identify correspondences of tags, which may eventually represent semantic meanings of tags.

Keywords microblog, user tag suggestion, tag correspondence model, probabilistic graphical model, context

1 Introduction

Microblogging is a broadcast medium in Web 2.0. Different from traditional blog services, microblogs allow users to exchange small elements of content such as short sentences, single images or video links. Due to the convenience of the production, spread and consumption of short messages, microblogging is growing into a popular platform for sharing information and expressing opinions. Microblog users generate rich contents including short messages and comments. Meanwhile, microblog users build a complex social network with following or forwarding behaviors. Both user generated content and social networks constitute the context information of a microblog user.

The nature of microblogs is to provide a new way of interaction for users. Therefore, it is crucial for microblog services to be able to recommend appropriate information that users are interested in. In order to well understand the interests of users, some microblog services encourage users to label tags to themselves. We take Kaifu Lee, a popular user on Sina Weibo (Sina Weibo is the largest microblog service in China, and in use by over 30% of Chinese Internet users. One can access via http://weibo.com.) for example. Kaifu Lee is the CEO of "Innovation Works", an IT company that aims to create successful Chinese start-ups in Internet and mobile Internet. He published his autobiography entitled with "Making a World of Difference". Hence, he annotates himself with the following tags:

Regular Paper
Special Section on Social Media Processing
This work is supported by the National Natural Science Foundation of China under Grant Nos. 61170196 and 61202140, and the Major Project of the National Social Science Foundation of China under Grant No. 13&ZD190.
A preliminary version of the paper was published in the Proceedings of SMP 2014.
*Corresponding Author

"venture investment", "microblog fans", "Innovation Works", "education", "technology", "e-business", "mobile internet", "start-ups", "Internet", and "Making a World of Difference". These tags provide a powerful scheme to represent attributes or interests of microblog users, and may eventually facilitate personalized recommendation and information retrieval.

User tags are all annotated independently by self, hence being noisy and disorganized. In order to profoundly understand user tags, it is intuitive to represent implicit semantics of user tags using correspondences identified from the rich context of microblog users. Here each correspondence is referred to as a unique element in the context which is semantically correlated with the tag. For example, for the tag "mobile internet" of Kaifu, we may identify the word "IT" in his self-description as a correspondence.

In general, the context information of microblog users origins from multiple sources. Each source has its own correspondence candidates. The sources can be categorized into two major types: user-oriented ones and neighbor-oriented ones. Here, B is the neighbor of A if A follows B in Sina Weibo.

User-Oriented Sources. The information generated by users themselves is defined as user-oriented sources, such as short messages and user profiles. These user generated contents usually reveal the interests and attributes of a user. It is thus probable to find correspondences of user tags from these sources.

Neighbor-Oriented Sources. The information from neighbor users of the given user is defined as neighbor-oriented sources, such as tags and short messages generated by these neighbor users. As the saying goes that "birds of a feather flock together", a user usually has common attributes or interests with its neighbor users. This has been verified in sociology[1]. Hence, it is feasible to identify correspondences from the neighbor-oriented sources for a user's tags.

To find precise correspondences of tags from these sources, two facts make it extremely challenging.

1) The context information is complex and noisy. For example, each user may generate many short messages with diverse topics and in informal styles, which makes it difficult to identify appropriate correspondences of tags.

2) The context information is from multiple and heterogenous sources, and each source has its own characteristics. It is non-trivial to jointly model multiple sources.

To address the challenges, we propose a probabilistic generative model, Tag Correspondence Model (TCM), to infer correspondences of user tags from multiple sources. For each source, we carefully select semantic elements as correspondence candidates. Take short messages for example, we can use either words or latent topics obtained from these messages as correspondence candidates. TCM will iteratively learn a probabilistic distribution over tags for each correspondence. TCM can also automatically adjust the proportion of correspondences from different sources with respect to the characteristics of each user.

It is straightforward for TCM to suggest tags for those users who have not annotated any tags according to their context information. For experiments, we build a real-world dataset and take user tag suggestion as our quantitative evaluation task. Experimental results show that TCM outperforms the state-of-the-art methods for microblog user tag suggestion, which indicates that TCM can efficiently identify correspondences of tags from the rich context information of users.

2 Related Work

There has been broad spectrum of studies on general social tag modeling and personalized social tag suggestion. These studies mostly focus on the tagging behaviors of a user on online items such as Web pages, images, and videos.

As a personalized recommendation task, some successful techniques in recommender systems are introduced to address the task of social tag suggestion, e.g., user/item based collaborative filtering[2-4], matrix and tensor decomposition[5-7]. Some graph-based methods are also explored for social tag suggestion[8]. In these methods, a tripartite user-item-tag graph is built based on the history of user tagging behaviors, and random walks are performed over the graph to rank tags. We categorize these methods into the collaboration-based approach.

The above mentioned studies on social tag suggestion are all based on the history of tagging behaviors. There are also many researches focusing on recommending tags based on meta-data of items, which are usually categorized into the content-based approach. Some researchers consider each social tag as a classification category, and thereby address social tag suggestion as a task of multi-label classification[9-14]. In these methods, the semantic relations between features and tags are implicitly hidden behind the parameters of classifiers, and thus are usually not human interpretable.

Inspired by the popularity of latent topic models such as Latent Dirichlet Allocation (LDA)[15], various graphical methods have been proposed to model the semantic relations of users, items, and tags for social tag suggestion. An intuitive idea is to consider both tags and words as being generated from the same set of latent topics. By representing both tags and descriptions as the distributions of latent topics, it suggests tags according to the likelihood given the meta-data of items[16-18]. As an extension, Bundschus *et al.*[19] proposed a joint latent topic model of users, words, and tags. Furthermore, an LDA-based topic model, Content Relevance Model (CRM)[20], was proposed to find the content-related tags for suggestion. Its experiments show the outperformance compared with both classification-based methods and Corr-LDA[21], a typical topic model for modeling both document contents and annotations.

Despite the importance of modeling microblog user tags, there has been little work focusing on this. Unlike other social tagging systems, in microblog user tagging systems, each user can only annotate tags to him/herself. Hence, we are not able to adopt the collaboration-based approach. Since we want to interpret semantic meanings of user tags, the classification-based methods are not competent either. Considering the powerful representation ability of graphical models, in this paper, we propose Tag Correspondence Model (TCM). Although some graphical models have been proposed for other social tagging systems as mentioned above, most of them are designed for modeling semantic relations between tags and some limited and specific factors, such as users or words, and thus are not capable of joint modeling of rich context information. On the contrary, TCM can identify complex and heterogeneous correspondences of user tags from multiple sources. In our experiments, we will show that it is by no means unnecessary to consider rich context for modeling microblog user tags.

3 Tag Correspondence Model

We give some formalized notations and definitions before introducing TCM. Suppose we have a collection of microblog users U. Each user $u \in U$ will generate rich text information such as self-description and short messages, annotate itself with a set of tags $\boldsymbol{a}_u$ from a vocabulary T of size $|T|$, and also build friendship with a collection of neighbor users $\boldsymbol{f}_u$.

3.1 The Model

We propose Tag Correspondence Model (TCM) to identify correspondences of each tag from multiple sources of users including but not limited to self-descriptions, short messages, and neighbor users. We design TCM as a probabilistic generative model.

We show the graphical model of TCM in Fig.1. In TCM, without loss of generality, we denote all sources of a user as a set S_u and all tags of a user as A_u. Each source $s \in S_u$ is represented as a weighted vector $\boldsymbol{x}_{u,s}$ over a vocabulary space V_s. All elements in these vocabularies are considered as correspondence candidates. Each correspondence r from the source s is represented as a multinomial distribution $\boldsymbol{\phi}_{s,r}$ over all tags in the vocabulary T drawn from a symmetric Dirichlet prior β. The annotated tags of a microblog user u is generated by first drawing a user-specific mixture $\boldsymbol{\pi}_u$ from asymmetric Dirichlet priors η_u, which indicates the distribution of each source for the user. For each source s, a user-specific mixture $\boldsymbol{\theta}_{u,s}$ over V_s correspondences is drawn from asymmetric Dirichlet priors $\boldsymbol{\alpha}_{u,s}$, which indicate the prior importance of correspondences for user. Suppose $\boldsymbol{x}_{u,s}$ indicates the normalized importance scores of all correspondences in source s for user u. We denote the prior of each correspondence r as $\alpha_{u,s,r} = \alpha x_{u,s,r}$, where α is the base score which can be manually pre-defined as in LDA[22].

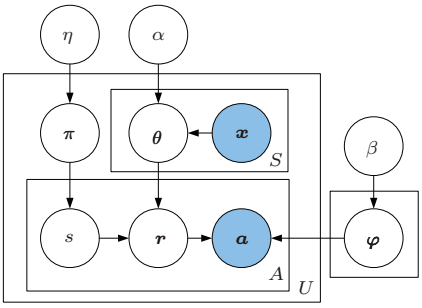


Fig.1. Tag Correspondence Model.

In TCM, the generative process of each tag t annotated by user u is shown as follows:

1) picking a source s from $\boldsymbol{\pi}_u$,

2) picking a correspondence r from $\boldsymbol{\theta}_{d,s}$, and

3) picking a tag t from $\boldsymbol{\phi}_{s,r}$.

Hence, tag t will be picked eventually in proportion to how much the user prefers source s, how much source s prefers correspondence r, and how much correspondence r prefers tag t.

Note that one of these sources will be interpreted as a global source, which contains only one correspondence and is available for each user. In this paper, we assume that each tag annotated by users can be explained by sources of themselves. But in fact, some popular tags are generally annotated that we cannot assign an appropriate correspondence to them. Thus we bring in the global source to overcome this situation. When an annotated tag cannot find an appropriate correspondence from other sources, it will be considered as being generated from the global correspondence.

In TCM, the annotated tags and the prior importance of correspondences in multiple sources are observed, and thus shaded in Fig.1. We are required to find an efficient way to measure the joint likelihood of observed tags $\boldsymbol{a}$ and unobserved source and correspondence assignments, i.e., $\boldsymbol{s}$ and $\boldsymbol{r}$, respectively. The joint likelihood is formalized as follows,

$$\Pr(\boldsymbol{a},\boldsymbol{s},\boldsymbol{r}|\boldsymbol{x},\alpha,\eta,\beta)=\prod_{u\in U}\Pr(\boldsymbol{a}_u,\boldsymbol{s}_u,\boldsymbol{r}_u|\boldsymbol{x}_u,\alpha,\eta,\beta).$$

Given a user u, we use $\boldsymbol{a}_u$, $\boldsymbol{s}_u$, $\boldsymbol{r}_u$ and $\boldsymbol{x}_u$ to represent the observed variables and correspondence assignments of u. We omit the subscript of vectors and formalize the right part as follows,

$$\Pr(\boldsymbol{a},\boldsymbol{s},\boldsymbol{r}|\boldsymbol{x},\alpha,\eta,\beta)=\Pr(\boldsymbol{a}|\boldsymbol{r},\beta)\Pr(\boldsymbol{r},\boldsymbol{s}|\boldsymbol{x},\alpha,\eta).$$

By optimizing the joint likelihood, we will derive the updates for parameters of TCM including $\boldsymbol{\pi}$, $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. In this joint likelihood, the first item $\Pr(\boldsymbol{a}|\boldsymbol{r},\beta)$ is similar to the word generation in LDA and thus we use the same derivation as in [22]. The second term can be decomposed as follows,

$$\Pr(\boldsymbol{r},\boldsymbol{s}|\boldsymbol{x},\alpha,\eta)=\Pr(\boldsymbol{r}|\boldsymbol{s},\boldsymbol{x},\alpha)\Pr(\boldsymbol{s}|\eta).$$

Following the equation (52) in [23], these two parts can be further formalized as

$$\begin{aligned}\Pr(\boldsymbol{s}|\boldsymbol{\eta}) &= \int_{\boldsymbol{\pi}}\Pr(\boldsymbol{s}|\boldsymbol{\pi})\Pr(\boldsymbol{\pi}|\boldsymbol{\eta})\mathrm{d}\boldsymbol{\pi}\\ &= \int_{\boldsymbol{\pi}}\prod_{i=1}^{|\boldsymbol{x}|}(Multi(s_i|\boldsymbol{\pi}))Dir(\boldsymbol{\pi}|\boldsymbol{\eta})\mathrm{d}\boldsymbol{\pi}\\ &= \frac{\Delta(n_{u,:,\cdot,\cdot}+\boldsymbol{\eta})}{\Delta(\boldsymbol{\eta})},\end{aligned}$$

and

$$\begin{aligned}&\Pr(\boldsymbol{r}|\boldsymbol{s},\boldsymbol{x},\boldsymbol{\alpha})\\ &= \int_{\boldsymbol{\theta}}\Pr(\boldsymbol{r}|\boldsymbol{\theta},\boldsymbol{s})\Pr(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{\alpha})\mathrm{d}\boldsymbol{\theta}\\ &= \int_{\boldsymbol{\theta}}\prod_{i=1}^{|\boldsymbol{x}|}(Multi(r_i|\theta_{s_i}))\prod_{i=1}^{|\boldsymbol{x}|}(Dir(\boldsymbol{\theta}_{x_i}|\boldsymbol{\alpha}))\mathrm{d}\boldsymbol{\theta}\\ &= \prod_{s_i}\frac{\Delta(n_{u,s_i,:,\cdot}+\boldsymbol{\alpha}_{u,s_i})}{\Delta(\boldsymbol{\alpha}_{u,s_i})}.\end{aligned}$$

Here, $\Delta(\boldsymbol{\alpha})$ is the "Dirichlet delta function"[23], $Multi$ indicates multinomial distribution and Dir indicates Dirichlet distribution. We denote the count $n_{u,j,k,t}$ as the number of occurrences of the source $j\in S_u$, the correspondence $k\in V_j$ as being assigned to the tag $t\in T$ of user u. We further sum counts using "$\cdot$" and select a vector of counts using ":".

We observe that each correspondence is only allocated in one source, and thus there is no need to explicitly use the sources $\boldsymbol{s}$. We can use Gibbs Sampling[24] to track the correspondence assignments $\boldsymbol{r}$. Following the derivations of LDA[22], the sampling update equation of assigning a new source and correspondence for a tag is formalized as follows,

$$\begin{aligned}&\Pr(s_{u,i}=j,r_{u,i}=k|\boldsymbol{s}_{\neg u,i},\boldsymbol{r}_{\neg u,i},a_{u,i}=t,\alpha,\beta,\eta)\\ &= \hat{p}_{(\neg u,i)}(a_{u,i}=t|r_{u,i}=k)\hat{p}_{(\neg u,i)}(r_{u,i}=k|s_{u,i}=j)\\ &\quad \hat{p}_{(\neg u,i)}(s_{u,i}=j),\end{aligned}$$

in which the three parts can be further formalized as

$$\hat{p}_{(\neg u,i)}(a_{u,i}=t|r_{u,i}=k)=\hat{\phi}_{s,r,t}=\frac{n^{(\neg u,i)}_{\cdot,j,k,t}+\beta}{n^{(\neg u,i)}_{\cdot,j,k,\cdot}+|T|\beta},$$

$$\hat{p}_{(\neg u,i)}(r_{u,i}=k|s_{u,i}=j)=\hat{\theta}_{u,s,r}=\frac{n^{(\neg u,i)}_{u,j,k,\cdot}+\alpha_{u,j,k}}{n^{(\neg u,i)}_{u,j,\cdot,\cdot}+\alpha_{u,j,\cdot}},$$

$$\hat{p}_{(\neg u,i)}(s_{u,i}=j)=\hat{\pi}_{u,s}=\frac{n^{(\neg u,i)}_{u,j,\cdot,\cdot}+(\boldsymbol{\alpha}_S)_j}{n^{(\neg u,i)}_{u,\cdot,\cdot,\cdot}+\sum_{j\in S}(\boldsymbol{\alpha}_S)_j}.$$

With above equations, we can get

$$\begin{aligned}&\Pr(s_{u,i}=j,r_{u,i}=k|\boldsymbol{s}_{\neg u,i},\boldsymbol{r}_{\neg u,i},a_{u,i}=t,\alpha,\beta,\eta)\\ &= \frac{n^{(\neg u,i)}_{\cdot,j,k,t}+\beta}{n^{(\neg u,i)}_{\cdot,j,k,\cdot}+|T|\beta}\times\frac{n^{(\neg u,i)}_{u,j,k,\cdot}+\alpha_{u,j,k}}{n^{(\neg u,i)}_{u,j,\cdot,\cdot}+\alpha_{u,j,\cdot}}\times\\ &\quad \frac{n^{(\neg u,i)}_{u,j,\cdot,\cdot}+(\boldsymbol{\alpha}_S)_j}{n^{(\neg u,i)}_{u,\cdot,\cdot,\cdot}+\sum_{j\in S}(\boldsymbol{\alpha}_S)_j}\\ &\propto \frac{n^{(\neg u,i)}_{\cdot,j,k,t}+\beta}{n^{(\neg u,i)}_{\cdot,j,k,\cdot}+|T|\beta}\big(n^{(\neg u,i)}_{u,j,k,\cdot}+\alpha x_{u,j,k}\big).\end{aligned}\tag{1}$$

Here the sign $\neg u,i$ indicates that the count excludes the current assignment. For simplicity, we also define $(\boldsymbol{\alpha}_S)_j=\alpha_{u,j,\cdot}$, and thus the numerator in the second fraction cancels the denominator in the last fraction. Moreover, the denominator in the second fraction is

constant for different source and correspondence assignments, and thus it is dropped in (1). We can observe that the update rule is quite similar to that of LDA.

For learning and inference, we can estimate the hidden parameters in TCM based on the collapsed sampling formula in (1). We can efficiently compute the counts n as the number of times that each tag has been assigned with each source and each correspondence. A sampler will iterate over the collection of users, reassign sources and correspondences, and update the counts. Finally, we can estimate the parameters of TCM using the source and correspondence assignments, in which we are mostly interested in

$$\pi_{u,s} = \frac{n_{u,s,\cdot,\cdot} + \eta}{n_{u,\cdot,\cdot,\cdot} + |S|\eta}, \tag{2}$$

$$\theta_{u,s,r} = \frac{n_{u,s,r,\cdot} + \alpha x_{u,s,r}}{n_{u,s,\cdot,\cdot} + \alpha x_{u,s,\cdot}}, \tag{3}$$

$$\phi_{s,r,t} = \frac{n_{\cdot,s,r,t} + \beta}{n_{\cdot,s,r,\cdot} + |T|\beta}. \tag{4}$$

3.2 Microblog User Tag Suggestion Using TCM

After obtaining the TCM model, the semantic meanings of a tag t can be represented using its correspondences, i.e., $\phi_{s,r,t} = \Pr(r|t)$, which can be used for further tag analysis such as clustering, classification, and suggestion. Here we introduce the method of microblog user tag suggestion using TCM.

Given a user u with sources $s \in S$ and correspondences $r \in V_s$, the probability of selecting a tag t is formalized as

$$\Pr(t|u,\phi) = \sum_{s\in S}\sum_{r\in V_s} \Pr(t|r,\phi)\Pr(r|u,s)\Pr(s|u),$$

where $\Pr(t|r,\phi) = \phi_{s,r,t}$, $\Pr(r|u,s) = \theta_{u,s,r}$, and $\Pr(s|u)$ indicates the preference of each source s given user u. Here we approximate $\Pr(s|u)$ using a global preference score of each source $\Pr(s)$, i.e., $\Pr(s|u) = \Pr(s)$. To compute $\Pr(s)$, we build a validation set to evaluate the suggestion performance with each source separately. By regarding the performance (e.g., F-measure when we suggest 10 tags) as the confidence to the source, we assign $\Pr(s)$ as the normalized evaluation score of s. Then, we rank all candidate tags in descending order and select top ranked tags for suggestion.

4 Selecting Sources and Correspondences

We introduce in detail each source with its correspondences that will be used in TCM. We also define weighting measures for correspondences of each source, which will be used as prior knowledge $\boldsymbol{x}$ in the joint likelihood.

4.1 User-Oriented Sources

In this paper, we consider the following two user-oriented sources: short messages and self-descriptions.

Short Messages. For short messages posted by a user, we have many choices of correspondence candidates, such as words and latent topics. In this paper we use words as correspondences. We measure the importance of each word in short messages of the user u according to its two statistical factors:

1) the ratio of short messages of u that contain the given word, named as message frequency;

2) the ratio of all users in U who have used the word, named as user frequency.

Inspired by term frequency and inverse document frequency (TF-IDF)[25], we define message frequency and inverse user frequency (MF-IUF) for each word w in short messages of user u, formalized as $MF\text{-}IUF_{u,w} = \frac{|M_{u,w}|}{|M_u|} \times \log\frac{|U|}{|U_w|}$, where $M_{u,w}$ is the set of messages that are posted by u and contain w, M_u is the set of all messages posted by u, and U_w is the set of all users in U that have used w in their short messages.

Self-Descriptions. A microblog user usually provides a short sentence for self-description. Although the description is short, usually with only tens of words, it contains dense information about the attributes or interests of the user. Similar to TF-IDF and MF-IUF, we define $UF\text{-}IUF_{u,w} = \frac{n_{u,w}}{n_{u,\cdot}} \times \log\frac{|U|}{|U_w|}$ for term weighting, where $n_{u,w}$ is the number of the times that u uses w in its self-description, and here U_w is the set of users in U who use w in their descriptions.

4.2 Neighbor-Oriented Sources

In this paper, we consider the following two neighbor-oriented sources: neighbor tags and neighbor descriptions.

We are aware that there are several methods incorporating network information into graphical models, such as Network Regularized Statistical Topic Model (NetSTM)[26] and Relational Topic Model (RTM)[27]. The basic idea of these methods is to smoothen the topic distribution of a document with its neighbor documents. Although these methods provide an effective approach to integrating both user-oriented and neighbor-oriented information, they suffer from two major issues.

1) These methods are not intuitively capable of modeling complex correspondences from multiple sources.

2) When modeling a document, the methods take its neighbor documents and their up-to-date topic distributions into consideration, which will be memory and computation consuming.

Here we use a simple and effective way to model neighbor-oriented sources, whose effectiveness and efficiency will be demonstrated in our experiments.

Neighbor Tags. For a user u, the tags annotated by its neighbors reflect the interests and attributes of u's ego-network, and hence are applicable to be selected as correspondence candidates of u's tags. We also consider two factors to measure the importance of neighbor tags:

1) the ratio of neighbor users who have annotated the tag;

2) the ratio of all users in U who have annotated the tag.

Also motivated by the idea of TF-IDF, we define neighbor frequency and inverse user frequency (NF-IUF) for measuring the importance of each neighbor tag, $NF\text{-}IUF_{u,t} = \frac{|N_{u,t}|}{|N_{u,\cdot}|} \times \log \frac{|U|}{|U_t|}$, where $N_{u,t}$ is the set of u's neighbor users who have annotated themselves with tag t, N_u is the set of u's neighbor users, and U_t is the set of users in U who have annotated themselves with tag t. The method will emphasize those tags that are locally frequently used by neighbor users of the given user u.

Neighbor Descriptions. Similar to neighbor tags, self-descriptions by neighbor users will also be an appropriate neighbor-oriented source. The weighting scheme also follows the idea of TF-IDF, defined as $NF\text{-}IUF_{u,w} = \frac{|N_{u,w}|}{|N_u|} \times \log \frac{|U|}{|U_w|}$, where $N_{u,w}$ is the set of u's neighbor users who use the word w in their descriptions, and U_w is the set of users in U who use w in their descriptions.

5 Experiments and Analysis

We select Sina Weibo as our research platform. We randomly crawled 2 million users from Sina Weibo ranging from January 2012 to December 2012. Since a large percent of users in Sina Weibo fill in their profiles at will, these users with incredible information are interferences for training an efficient model. Thus we make another selection according to the quality of their information. From the raw data, we select 341 353 users, each having complete profiles, short messages, social networks and more than two tags. We also select 4 126 tags and each occurs more than 500 times. According to our statistics, the probability that these tags are annotated is as high as 98.67%. On average, each user has 4.54 tags, 63.35 neighbors and 305.24 neighbor tags, and each user description has 6.93 words.

In TCM, we set $\beta = 0.1$ following the common practice in LDA[22] and set $\alpha = 10$ so as to leverage the prior knowledge of correspondence candidates.

In experiments, we use UM, UD, NT and ND to stand for the following four sources: user messages, user descriptions, neighbor tags, and neighbor descriptions respectively.

In order to intuitively demonstrate the efficiency and effectiveness of TCM, in Subsection 5.1, we perform empirical analysis of learning results, including learning convergence, characteristic tags and correspondences of TCM. Then in Subsection 5.2, we perform quantitative evaluation on TCM by taking user tag suggestion as the target application.

5.1 Empirical Analysis

5.1.1 Learning Convergence

Although TCM is a bit more complex than plain LDA, it converges fast due to the incorporation of prior knowledge of each source. Fig.2 demonstrates the convergence trend of log-likelihood when training the models. The log-likelihood is computed over a small test set U_T using the learned TCM model after each iteration as

$$L(U_T) = \sum_{t \in U_T} \log \sum_{c,r} \Pr(t|c,r,\phi) \Pr(c,r|u).$$

We can observe that the log-likelihood starts to become stable around the 15th iteration. The convergence rate is fast according to the common practice in latent topic models[15]. This indicates the efficiency of TCM learning.

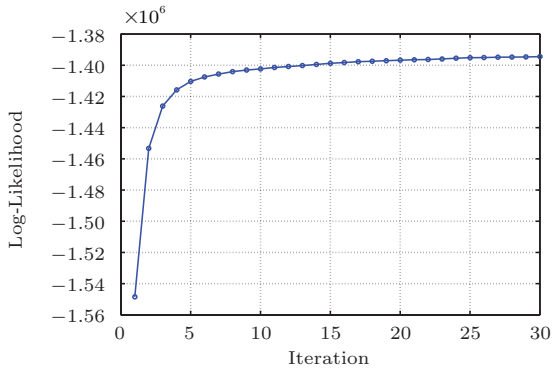


Fig.2. Convergence of learning process.

5.1.2 Characteristic Tags of Sources

In order to better understand the four sources in Table 1, we show the ratio of each source $\Pr(s)$ and top 5 characteristic tags assigned to various sources. Here $\Pr(s)$ is computed by simply aggregating all source assignments for tags in U, i.e.,

$$\Pr(s) = \frac{n_{\cdot,s,\cdot,\cdot} + \eta}{n_{\cdot,\cdot,\cdot,\cdot} + |S|\eta}.$$

We select representative tags of each source according to their characteristic scores in the source. Following the idea in [28], the characteristic score of tag t in source s is defined as

$$C(s,t) = \Pr(t|s) \times \Pr(s|t),$$

where

$$\Pr(t|s) = \frac{n_{\cdot,s,\cdot,t} + \beta}{n_{\cdot,s,\cdot,\cdot} + |T|\beta},$$
$$\Pr(s|t) = \frac{n_{\cdot,s,\cdot,t} + \beta}{n_{\cdot,\cdot,\cdot,t} + |S|\beta}.$$

To facilitate understanding, we explain some confusing tags in Table 1 as follows. "Fang Datong" is a Chinese popstar. Chongqing, Shenzhen and Guangzhou are large cities in China. In the tag "Taobao Shopkeeper", Taobao is a popular C2C service. "Douban" is a book review service in China.

Table 1. Proportion of Each Source and Its Characteristic Tags

Source	$\Pr(s)$	Top 5 Characteristic Tags
UM	0.19	Mobile internet, Fang Datong, Chongqing, Shenzhen, Guangzhou
UD	0.19	Plane model, Taobao Shopkeeper, photographer, cosplay, e-business
NT	0.42	Online shopping, novel, medium, reading, advertising
ND	0.20	Douban, lazy, novel, food, music

Note: UM, UD, NT, and ND stand for the following four sources: user messages, user descriptions, neighbor tags, and neighbor descriptions respectively.

From the statistics in Table 1, we can see that neighbor-oriented sources are more important than user-oriented sources. What is more, the source of neighbor tags occupies the most important place in the four sources with a ratio of 0.42. The superiority of neighbor-oriented sources is not surprised. A user generates user-oriented content all by himself/herself with much discretionary subjectivity, and thus may not necessarily fully reflect the corresponding user tags. Meanwhile, tags and descriptions of neighbors can be regarded, to some extent, as collaborative annotations to this user from his/her many friends, and thus may be more reasonable and less noisy.

Another observation from Table 1 is that the most characteristic tags of neighbor-oriented sources reflect the interests of users, such as "online shopping", "reading", "food" and "music". On the contrary, most characteristic tags of user-oriented sources uncover the attributes of users, such as occupations, locations, and identities. This indicates that attribute tags may tend to find good correspondences from user-oriented sources, and meanwhile interest tags from neighbor-oriented sources.

Note that the setting of global source in TCM is important for modeling user tags. The global source collects the tags with no appropriate correspondences. The top 5 tags assigned to the global source are "music", "movie", "food", "80s" and "travel". These tags are usually general and popular, and have less correlation with the context information of users. If there is no global source, these tags will annoy the process of correspondence identification for other tags.

5.1.3 Characteristic Correspondences of Tags

The mission of TCM is to find appropriate correspondences for user tags. Here we pick some tags annotated by Kaifu Lee as examples. In Table 2, we list characteristic correspondences of these tags. The characteristic score of correspondence r with tag t is computed as $C(r,t) = \Pr(t|r) \times \Pr(r|t)$. After each correspondence, we provide the source in brackets. From these tags and their correspondences, it is convinced that TCM can identify appropriate correspondences from noisy and heterogeneous sources.

Table 2. Characteristic Correspondences of Kaifu's Tags

Tag	Top-5 Characteristic Correspondence
Education	Internet (NT), education (UD), education (UM), politics (NT), study (NT)
Technology	Android (NT), Internet (NT), product (ND), create (ND), communication (NT)
Start-ups	Start-ups (NT), venture capital (NT), e-business (NT), entrepreneur (NT), Internet (UD)
Mobile internet	SNS (NT), mobile (UD), Internet (UM), mobile (UM), IT (NT)
E-business	B2C (NT), IT (NT), e-business (UM), e-business (NT), marketing (NT)

5.2 Evaluation on User Tag Suggestion

5.2.1 Evaluation Metrics and Baseline Methods

For the task of microblog user tag suggestion, we use precision, recall, and F-measure for evaluation. Given a microblog user, we denote its annotated tags (gold standard) as T_a, the suggested tags as T_s, and the correctly suggested tags as $T_s \cap T_a$. Then its precision, recall, and F-measure are defined as

$$P = \frac{T_s \cap T_a}{T_s}, R = \frac{T_s \cap T_a}{T_a}, F = \frac{2PR}{P+R}.$$

We perform 5-fold cross validation for each method, and use the averaged precision, recall, and F-measure over all test instances for evaluation. In experiments, the number of suggested tags M ranges from 1 to 10.

For microblog user tag suggestion, we select kNN[25], TagLDA[17], and NetSTM[26] as baseline methods for comparison. kNN is a typical classification algorithm based on closest training examples. TagLDA is a representative method of latent topic models for which one can refer to [17] for detailed information. In this paper, we modify original NetSTM[26] by regarding tags as explicit topics, which can thus model the semantic relations between user-oriented contents with tags and take the neighbor tag distributions for smoothing. We set the number of topics $K = 200$ for TagLDA, the number of neighbors $k = 5$ for kNN, and the regularization factor $\lambda = 0.15$ for NetSTM, by which they obtain the best performance.

5.2.2 Comparison Results

In Fig.3, we show the precision-recall curves of different methods for microblog user tag suggestion. Here we use TCM-XX to indicate the method of TCM with different sets of sources indicated by XX, which can be UM, UD, NT or ND. Moreover, TCM-UN indicates the combination of both user-oriented and neighbor-oriented sources.

In Fig.3, each point of a precision-recall curve represents suggesting different numbers of tags from $M = 1$ to $M = 10$. The point of $M = 1$ is at bottom right with higher precision but lower recall, while the point of $M = 10$ is at upper left with higher recall but lower precision. The closer a curve is to the upper right, the better the overall performance of the corresponding method will be.

From Fig.3, we observe that TCM significantly outperforms other baseline methods consistently except when it uses only short messages of users as the correspondence source. This indicates that the source of short messages in isolation is too noisy to suggest good user tags. We also find that TCM-UN achieves the best performance. When the suggestion number is $M = 10$, the F-measure of TCM-UN is 0.184 while that of the best baseline method NetSTM is 0.142. This verifies the necessity of joint modeling of multiple sources for user tag suggestion.

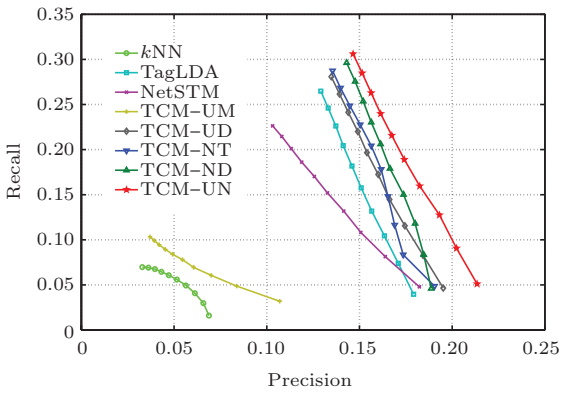


Fig.3. Evaluation results of different methods.

In three baseline methods, kNN and Tag-LDA only consider the user-oriented source (i.e., self-descriptions). The poor performance of kNN is not surprising because self-descriptions are usually too short to compute appropriate user similarities. Although NetSTM models more sources with both neighbor tags and user descriptions, it goes behind Tag-LDA when suggesting more tags. This indicates that it is non-trivial to fuse multiple sources for user tag suggestion.

Note that from Fig.3, we find that the absolute evaluation scores of the best method TCM-UN are low compared with other social tagging systems[6,17]. This is mainly caused by the characteristics of microblog user tagging systems. On one side, since each user can only be annotated by himself/herself, the annotated tags will be more arbitrary compared with other social tagging systems which are usually annotated collaboratively by thousands of users. On the other side, we perform evaluation by strictly matching suggested tags with user annotated tags. Hence, even a method can suggest reasonable tags for a user, which may usually have not been annotated by the specific user. Therefore, the evaluation scores can be used for comparing performance among methods, but are not applicable for judging the real performance of a method.

We also investigate the overlapping ratios of tags correctly suggested by different sources, as shown in Table 3. For each source of a line, the second column is

the number of tags correctly suggested by the source, and the third to the sixth columns record the ratios of common correct tags in this source. We find that the overlapping ratios are generally low, most of which are lower than 50%. This further verifies the need of joint modeling of multiple sources for user tag suggestion.

Table 3. Overlapping Ratios of Tags

Source	Number of Correctly Suggested Tags	UM	UD	NT	ND
UM	12 707	-	0.517	0.481	0.428
UD	16 191	0.406	-	0.593	0.403
NT	19 856	0.308	0.484	-	0.292
ND	16 038	0.339	0.407	0.362	-

5.2.3 Case Study

In Table 4, we show the top 5 tags suggested by TCM using various sources for the user Kaifu Lee we mentioned in Section 1. By taking the annotations of Kaifu as standard answers, we can see that most suggested tags are correct. What is more, although some suggested tags such as "Google", "marketing", "travel", "movie", and "reading" are not actually annotated by Kaifu, these tags are, to some extent, relevant to Kaifu according to his context. This also suggests that even though the absolute evaluation scores of user tag suggestion are lower compared with some other research tasks, it does not indicate poor performance, but is caused by the strategy of complete matching with user annotations in evaluation.

Table 4. Tags Suggested to Kaifu Lee from Different Sources

	Top 5 Suggested Tags
UM	**Mobile internet**, **start-ups**, **Internet**, **e-business**, indoors-man
UD	**Innovation**, freedom, **Internet**, Google, **start-ups**
NT	**Internet**, movie, **start-ups**, travel, **e-business**
ND	**Internet**, **start-ups**, **e-business**, marketing, **mobile internet**
UN	**Start-ups**, **e-business**, **Internet**, **mobile internet**, reading

Note: tags in bold letters are correct ones.

6 Conclusions and Future Work

In this paper, we formalized the task of modeling microblog user tags. We proposed a probabilistic generative model, TCM, to identify correspondences as a semantic representation of user tags. In TCM, we investigated user-oriented and neighbor-oriented sources for modeling. We carried out experiments on a real-world dataset, and the results showed that TCM can effectively identify correspondences of user tags from rich context information. Moreover, as a solution to microblog user tag suggestion, TCM achieves the best performance compared with baseline methods. Though we adopted the user tag suggestion task in Sina Weibo for evaluation, TCM is not application-specific. It can be easily extended to many other scenarios. For example, many online items in social tag suggestion, such as images, books, and videos, also suffer from the issue of rich context, which can thus benefit from TCM.

We will explore the following directions as future work. 1) In this paper, we perform strict matching between suggested tags and user annotated tags for evaluation, which makes the evaluation scores cannot reflect the real performance of tagging methods. We will investigate more reasonable evaluation methods, such as crowdsourcing, to better quantitatively measure the real performance of tagging methods. 2) We will explore more rich sources to improve the performance of microblog user tag suggestion. 3) We will explore user factors for measuring $\Pr(s|u)$ when suggesting tags with TCM as shown in Subsection 3.2.

References

[1] McPherson M, Smith-Lovin L, Cook J M. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 2001, 27: 415-444.

[2] Liang H, Xu Y, Li Y, Nayak R, Tao X. Connecting users and items with weighted tags for personalized item recommendations. In *Proc. the 21st ACM Conference on Hypertext and Hypermedia*, June 2010, pp.51-60.

[3] Peng J, Zeng D, Zhao H, Wang F. Collaborative filtering in social tagging systems based on joint item-tag recommendations. In *Proc. the 19th ACM International Conference on Information and Knowledge Management*, Oct. 2010, pp.809-818.

[4] Zhen Y, Li W, Yeung D. TagiCoFi: Tag informed collaborative filtering. In *Proc. the 3rd ACM Conference on Recommender Systems*, Oct. 2009, pp.69-76.

[5] Symeonidis P, Nanopoulos A, Manolopoulos Y. Tag recommendations based on tensor dimensionality reduction. In *Proc. the 2008 ACM Conference on Recommender Systems*, Oct. 2008, pp.43-50.

[6] Rendle S, Marinho L B, Nanopoulos A, Schmidt-Thieme L. Learning optimal ranking with tensor factorization for tag recommendation. In *Proc. the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, June 28-July 1, 2009, pp.727-736.

[7] Rendle S, Schmidt-Thieme L. Pairwise interaction tensor factorization for personalized tag recommendation. In *Proc. the 3rd ACM International Conference on Web Search and Data Mining*, Feb. 2010, pp.81-90.

[8] Jäschke R, Marinho L B, Hotho A, Schmidt-Thieme L, Stumme G. Tag recommendations in social bookmarking systems. *AI Communications*, 2008, 21(4): 231-247.

[9] Ohkura T, Kiyota Y, Nakagawa H. Browsing system for weblog articles based on automated folksonomy. In *Proc. the 15th International Conference on World Wide Web*, May 2006.

[10] Mishne G. AutoTag: A collaborative approach to automated tag assignment for weblog posts. In *Proc. the 15th International Conference on World Wide Web*, May 2006, pp.953-954.

[11] Lee S, Chun A. Automatic tag recommendation for the Web 2.0 blogosphere using collaborative tagging and hybrid ANN semantic structures. In *Proc. the 6th WSEAS International Conference on Applied Computer Science*, Apr. 2007, pp.88-93.

[12] Katakis I, Tsoumakas G, Vlahavas I. Multilabel text classification for automated tag suggestion. In *Proc. the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, volume 18, Sept. 2008.

[13] Fujimura S, Fujimura K, Okuda H. Blogosonomy: Autotagging any text using bloggers' knowledge. In *Proc. IEEE/WIC/ACM International Conference on Web Intelligence*, Nov. 2007, pp.205-212.

[14] Heymann P, Ramage D, Garcia-Molina H. Social tag prediction. In *Proc. the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, July 2008, pp.531-538.

[15] Blei D, Ng A, Jordan M. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 2003, 3: 993-1022.

[16] Krestel R, Fankhauser P, Nejdl W. Latent Dirichlet allocation for tag recommendation. In *Proc. the 3rd ACM Conference on Recommender Systems*, Oct. 2009, pp.61-68.

[17] Si X, Sun M. Tag-LDA for scalable real-time tag recommendation. *Journal of Computational Information Systems*, 2009, 6(1): 23-31.

[18] Liu Z, Tu C, Sun M. Tag dispatch model with social network regularization for microblog user tag suggestion. In *Proc. the 24th International Conference on Computational Linguistics*, Dec. 2012, pp.755-764.

[19] Bundschus M, Yu S, Tresp V, Rettinger A, Dejori M, Kriegel H. Hierarchical Bayesian models for collaborative tagging systems. In *Proc. the 9th IEEE International Conference on Data Mining*, Dec. 2009, pp.728-733.

[20] Iwata T, Yamada T, Ueda N. Modeling social annotation data with content relevance using a topic model. In *Proc. the 23rd Annual Conference on Neural Information Processing Systems*, Dec. 2009, pp.835-843.

[21] Blei D, Jordan M. Modeling annotated data. In *Proc. the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, July 28-August 1, 2003, pp.127-134.

[22] Griffiths T, Steyvers M. Finding scientific topics. *Proc. the National Academy of Sciences of the United States of America*, 2004, 101(Suppl 1): 5228-5235.

[23] Heinrich G. Parameter estimation for text analysis. Technical Report, vsonix GmbH + University of Leipzig, Germany, May 2005.

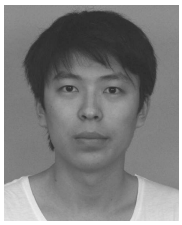
[24] Andrieu C, de Freitas N, Doucet A, Jordan M. An introduction to MCMC for machine learning. *Machine Learning*, 2003, 50(1/2): 5-43.

[25] Manning C D, Raghavan P, Schütze H. Introduction to Information Retrieval, Volume 1. Cambridge University Press, Cambridge, 2008.

[26] Mei Q, Cai D, Zhang D, Zhai C. Topic modeling with network regularization. In *Proc. the 17th International Conference on World Wide Web*, Apr. 2008, pp.101-110.

[27] Chang J, Blei D M. Relational topic models for document networks. In *Proc. the 12th International Conference on Artificial Intelligence and Statistics*, Apr. 2009, pp.81-88.

[28] Cohn D, Chang H. Learning to probabilistically identify authoritative documents. In *Proc. ICML*, June 29-July 2, 2000, pp.167-174.



Cun-Chao Tu is a Ph.D. student of the Department of Computer Science and Technology, Tsinghua University, Beijing. He got his B.E. degree in computer science from Tsinghua University in 2013. His research interests are user representation and social computation.

Zhi-Yuan Liu is an assistant researcher of the Department of Computer Science and Technology, Tsinghua University, Beijing. He got his B.E. and Ph.D. degrees in computer science from Tsinghua University in 2006 and 2011 respectively. His research interests are natural language processing and social computation. He has published over 40 papers in international journals and conferences including ACM Transactions, IJCAI, AAAI, ACL, and EMNLP. He was awarded Tsinghua Excellent Doctoral Dissertation in 2011, Excellent Doctoral Dissertation by Chinese Association for Artificial Intelligence in 2012, and Excellent Post-Doctoral Fellow Award at Tsinghua University in 2013.

Mao-Song Sun is a professor of the Department of Computer Science and Technology, Tsinghua University, Beijing. He got his B.E. and M.E. degrees in computer science from Tsinghua University in 1986 and 1988 respectively, and Ph.D. degree from the Department of Chinese, Translation and Linguistics, City University of Hong Kong, in 2004. His research interests include natural language processing, Chinese computing, Web intelligence, and computational social sciences. He has published over 150 papers in academic journals and international conferences in the above fields. He serves as a vice president of the Chinese Information Processing Society, the council member of CCF, the director of Massive Online Education Research Center of Tsinghua University, and the editor-in-chief of the Journal of Chinese Information Processing.