Inferring Correspondences from Multiple Sources for Microblog User Tags

Cunchao Tu, Zhiyuan Liu, and Maosong Sun

Department of Computer Science and Technology
State Key Lab on Intelligent Technology and Systems
National Lab for Information Science and Technology
Tsinghua University, Beijing 100084, China
{tucunchao,lzy.thu}@gmail.com, sms@tsinghua.edu.cn

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. 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 the 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: User Tag Suggestion, Tag Correspondence Model, Probabilistic Graphical Model.

1 Introduction

As microblogs grow in popularity, Microblog users generate rich contents everyday, which include 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. In order to well understand the interests of users, some microblog services encourage users to label tags to themselves. Tags provide a powerful scheme to represent attributes or interests of microblog users, and may eventually facilitate personalized recommendation and information retrieval.

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 a unique element in

H.-Y. Huang et al. (Eds.): SMP 2014, CCIS 489, pp. 1–12, 2014.

the context which is semantically correlated with the tag. For example, for the tag "mobile_internet" of Kai-Fu, we may identify the word "mobile" 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**.

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. Meanwhile, TCM can 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. Experiment 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. Many studies have been done to suggest tags for products such as books, movies and restaurants[8,15,7,17,10]. 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 [14], matrix and tensor decomposition [18]. 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**. For example, some researchers consider each social tag as a classification category, and thereby address social tag suggestion as a task of multi-label classification [9]. 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) [2], various graphical methods are 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 [16,11]. As an extension, [3] propose a joint latent topic model of users, words and tags. Furthermore, an LDA-based topic model, Content Relevance Model (CRM) [7], is proposed to find the content-related tags for suggestion, whose experiments show the outperformance compared to both classification-based methods and Corr-LDA [1], a typical topic model for modeling both 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 itself. 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 neither. 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. 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 $\phi_{s,r}$

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 π_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 $\theta_{u,s}$ over V_s correspondences is drawn from asymmetric Dirichlet priors $\boldsymbol{\alpha}_{u,s}$, which indicate the prior importance of correspondences for the user. Suppose $\boldsymbol{x}_{u,s}$ indicates the normalized importance scores of all correspondences in the source s for the 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 [6].

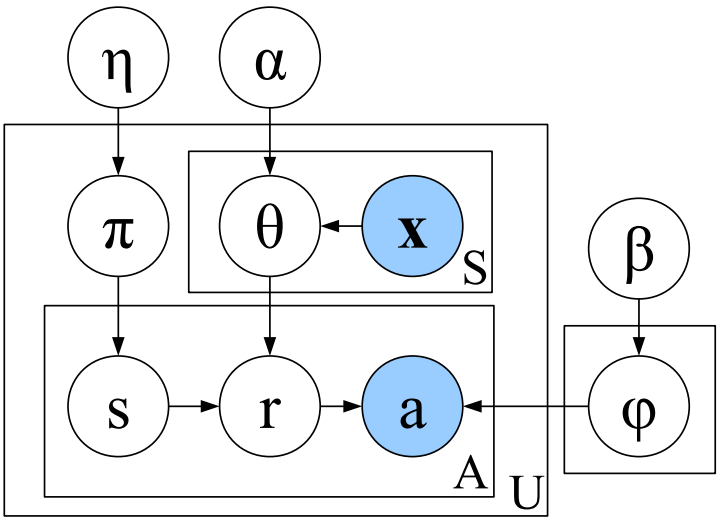

Fig. 1. Tag Correspondence Model

In TCM, the generative process of each tag t annotated by the user u is shown as follows: (1) picking a source s from π_u, (2) picking a correspondence r from $\theta_{d,s}$, and (3) picking a tag t from $\phi_{s,r}$. Hence, the tag t will be picked eventually in proportion to how much the user prefers the source s, how much the source s prefers the correspondence r, and how much the correspondence r prefers the tag t.

Note that one of these sources will be interpreted as a *global* source, which contains only one correspondence and presents on each user. When an annotated tag cannot find an appropriate correspondence from other sources, it will be considered as generated from the global correspondence.

In TCM, the annotated tags and the prior importance of correspondences in multiple sources are observed, which are thus shaded in Figure 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,

$$\Pr(\boldsymbol{a}, \boldsymbol{s}, \boldsymbol{r}|\boldsymbol{x}, \alpha, \eta, \beta) = \Pr(\boldsymbol{a}|\boldsymbol{r}, \beta) \Pr(\boldsymbol{r}, \boldsymbol{s}|\boldsymbol{x}, \alpha, \eta). \tag{1}$$

By optimizing the joint likelihood, we will derive the updates for parameters of TCM including π, θ and ϕ. 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 [6]. The second term can be decomposed as follows,

$$\Pr(\boldsymbol{r}, \boldsymbol{s}|\boldsymbol{x}, \alpha, \eta) = \Pr(\boldsymbol{r}|\boldsymbol{s}, \boldsymbol{x}, \alpha) \Pr(\boldsymbol{s}|\eta), \tag{2}$$

in which the two parts can be further formalized as

$$\Pr(\boldsymbol{s}|\eta) = \int_{\pi} \Pr(\boldsymbol{s}|\pi) \Pr(\pi|\eta) d\pi = \prod_{u \in U} \frac{\Delta(n_{u,:,\cdot,\cdot} + \boldsymbol{\eta})}{\Delta(\boldsymbol{\eta})}, \tag{3}$$

$$\Pr(\boldsymbol{r}|\boldsymbol{s}, \boldsymbol{x}, \alpha) = \int_{\theta} \Pr(\boldsymbol{r}|\theta, \boldsymbol{s}) \Pr(\theta|\boldsymbol{x}, \alpha) d\theta = \prod_{u \in U} \prod_{s \in S} \frac{\Delta(n_{u,s,:,\cdot} + \alpha_{u,s})}{\Delta(\boldsymbol{\alpha}_{u,s})}. \tag{4}$$

Here 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 the 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 to track the correspondence assignments $\boldsymbol{r}$. Following the derivations of LDA [6], the sampling update equation of assigning a new source and correspondence for a tag is formalized as follows,

$$\begin{aligned} &\Pr(s_{u,i} = j, r_{u,i} = k|\boldsymbol{s}_{\neg u,i}, \boldsymbol{r}_{\neg u,i}, a_{u,i} = t, \alpha, \beta, \eta) \\ &= \frac{n^{(\neg u,i)}_{\cdot,j,k,t} + \beta}{n^{(\neg u,i)}_{\cdot,j,k,\cdot} + |T|\beta} \cdot \frac{n^{(\neg u,i)}_{u,j,\cdot,\cdot} + (\boldsymbol{\alpha}_S)_j}{n^{(\neg u,i)}_{u,\cdot,\cdot,\cdot} + \sum_{j \in S} (\boldsymbol{\alpha}_S)_j} \cdot \frac{n^{(\neg u,i)}_{u,j,k,\cdot} + \alpha_{u,j,k}}{n^{(\neg u,i)}_{u,j,\cdot,\cdot} + \alpha_{u,j,\cdot}} \\ &\propto \frac{n^{(\neg u,i)}_{\cdot,j,k,t} + \beta}{n^{(\neg u,i)}_{\cdot,j,k,\cdot} + |T|\beta} \cdot \left(n^{(\neg u,i)}_{u,j,k,\cdot} + \alpha x_{u,j,k}\right). \end{aligned} \tag{5}$$

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 assignment, and thus it is dropped in the last formula. 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 Eq.(5). 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

$$\pi_{u,s} = \frac{n_{u,s,\cdot,\cdot} + \eta}{n_{u,\cdot,\cdot,\cdot} + |S|\eta} \tag{6}$$

$$\theta_{u,s,r} = \frac{n_{u,s,r,\cdot} + \alpha x_{u,s,r}}{n_{u,s,\cdot,\cdot} + \alpha x_{u,s,\cdot}} \tag{7}$$

$$\phi_{s,r,t} = \frac{n_{\cdot,s,r,t} + \beta}{n_{\cdot,s,r,\cdot} + |T|\beta}. \tag{8}$$

3.2 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

$$\Pr(t|u,\phi) = \sum_{s \in S} \sum_{r \in V_s} \Pr(t|r,\phi) \Pr(s,r|u) \Pr(s|u), \tag{9}$$

where $\Pr(s,r|u) = \Pr(r|u) = x_{u,s,r}$, and $\Pr(t|r,\phi) = \phi_{s,r,t}$. $\Pr(s|u)$ indicates the preference of each source s given the 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 at $M = 10$ in this paper) 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 Selection of 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 Equation (5). In this paper, we consider two user-oriented sources: user messages(UM) and user descriptions(UD). We also consider two neighbor-oriented sources: neighbor tags(NT) and neighbor descriptions(ND). Inspired by term frequency and inverse document frequency (TF-IDF), we define some similar ways to measure the importance of each candidate in each source.

There are several methods incorporating network information into graphical models, such as Network Regularized Statistical Topic Model (NetSTM) [13] and Relational Topic Model (RTM) [4]. The basic idea of these methods is to smooth the topic distribution of a document with its neighbor documents. Although these methods provide an effective approach to intergrading 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.

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. From the raw data, we select 341,353 users with each having complete profiles, short messages, social networks and more than 2 tags. In this dataset, the vocabulary size of tags is 4,126. 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 [6] 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.

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

5.1 Empirical Analysis

Characteristic Tags of Sources. In order to better understand the four sources, in Table 1[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.

$$\Pr(s) = \frac{n_{\cdot,s,\cdot,\cdot} + \eta}{n_{\cdot,\cdot,\cdot,\cdot} + |S|\eta}. \tag{10}$$

We select representative tags of each source according to their characteristic scores in the source. Following the idea in [5], the characteristic score of a tag t in a source s is defined as

$$C(s,t) = \Pr(t|s) \times \Pr(s|t), \tag{11}$$

[1] To facilitate understanding, we explain some confusing tags as follows. "Fang Datong" is a Chinese pop-star. 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.

where

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

Table 1. Proportion of each source and its characteristic tags. UM, UD, NT and ND stand for the following four sources, user messages, user descriptions, neighbor tags and neighbor descriptions.

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

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 itself 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 their many friends, and thus may be more reasonable and less noisy.

Another observation from Table 1 is that, the characteristic tags of neighbor-oriented sources most 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, 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. 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 correlations with the context information of users. If there is no global source, these tags will annoy the process of correspondence identification for other tags.

Characteristic Correspondences of Tags. The mission of TCM is to find appropriate correspondences for user tags. Here we pick some tags annotated

Table 2. Characteristic correspondences of Kai-Fu's tags

Tag	Top 5 Characteristic Correspondences
education	Internet (NT), eduction (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)

by Kai-Fu Lee as examples. In Table 2, we list characteristic correspondences of these tags. The characteristic score of a correspondence r with a 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.

5.2 Evaluation on Microblog User Tag Suggestion

Evaluation Metrics and Baseline Methods. For the task of microblog user tag suggestion, we use precision, recall and F-Measure for evaluation. We also perform 5-fold cross validation for each method, and use the averaged precision, recall and F-Measure over all test instances for evaluation.

For microblog user tag suggestion, we select kNN [12], TagLDA [16], and NetSTM [13] 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 [16] for detailed information. In this paper, we modify original NetSTM [13] 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 obtains best performance.

Comparison Results. In Fig. 2 we show the precision-recall curves of different methods for microblog user tag suggestion. Here we use TCM-XX to indicate the method TCM using different sources, where UN indicates the combination of both user-oriented and neighbor-oriented sources. Each point of a precision-recall curve represents suggesting different number of tags from $M = 1$ to $M = 10$.

From Fig. 2 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.

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 computing 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.

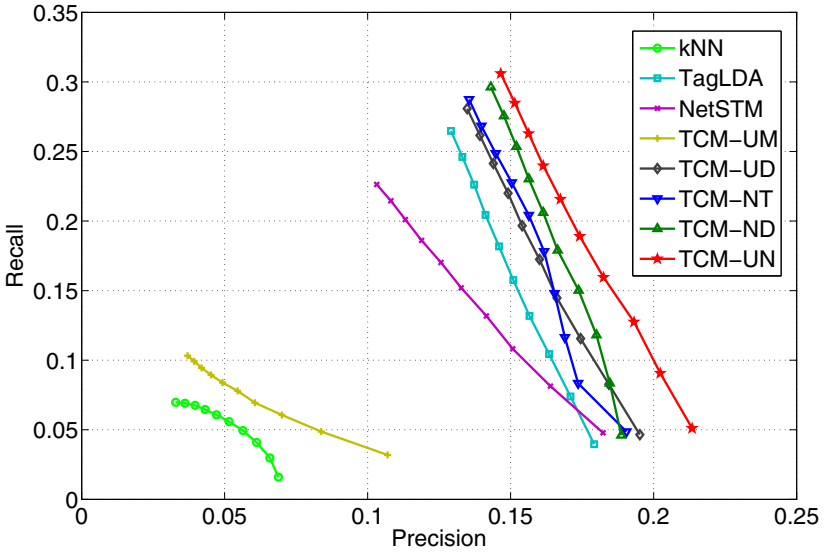

Fig. 2. Evaluation results of different methods

Note that, from Fig. 2 we find that the absolute evaluation scores of the best method TCM-UN are low compared with other social tagging systems [16,15]. This is mainly caused by the characteristics of microblog user tagging systems. On one side, since each user can only be annotated by itself, the annotated tags will be more arbitrary compared to 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.

Case Study. In Table 3 we show top 5 tags suggested by TCM using various sources for the user Kai-Fu Lee we mentioned in Section 1. By taking the annotations of Kai-Fu 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 Kai-Fu, these tags are, to some extent, relevant to Kai-Fu according to his context. This also suggests that, even though the absolute evaluation scores of user tag suggestion are lower compared to 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 3. Tags suggested to Kai-Fu 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

6 Conclusion and Future Work

In this paper, we formalize the task of modeling microblog user tags. We propose a probabilistic generative model, TCM, to identify correspondences as a semantic representation of user tags. In TCM we investigate user-oriented and neighbor-oriented sources for modeling, and carry out experiments on a real world dataset. The results show 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 to baseline methods.

We will explore the following directions as future work. (1) We will explore more rich sources to improve the performance of microblog user tag suggestion. (2) We will explore user factors for measuring $\Pr(s|u)$ when suggesting tags with TCM as shown in Section 3.2.

Acknowledgement. This work is supported by the National Natural Science Foundation of China (NSFC) under the grant No. 61170196 and 61202140.

References

1. Blei, D., Jordan, M.: Modeling annotated data. In: Proceedings of SIGIR, pp. 127–134. ACM (2003)
2. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. JMLR 3, 993–1022 (2003)
3. Bundschus, M., Yu, S., Tresp, V., Rettinger, A., Dejori, M., Kriegel, H.: Hierarchical bayesian models for collaborative tagging systems. In: Proceedings of ICDM, pp. 728–733 (2009)

4. Chang, J., Blei, D.M.: Relational topic models for document networks. In: Proceedings of AISTATS, pp. 81–88 (2009)
5. Cohn, D., Chang, H.: Learning to probabilistically identify authoritative documents. In: Proceedings of ICML, pp. 167–174 (2000)
6. Griffiths, T., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America 101(suppl. 1), 5228–5235 (2004)
7. Iwata, T., Yamada, T., Ueda, N.: Modeling social annotation data with content relevance using a topic model. In: Proceedings of NIPS, pp. 835–843 (2009)
8. Jaschke, R., Marinho, L., Hotho, A., Schmidt-Thieme, L., Stumme, G.: Tag recommendations in social bookmarking systems. AI Communications 21(4), 231–247 (2008)
9. Katakis, I., Tsoumakas, G., Vlahavas, I.: Multilabel text classification for automated tag suggestion. In: ECML PKDD Discovery Challenge 2008, p. 75 (2008)
10. Liu, Z., Chen, X., Sun, M.: A simple word trigger method for social tag suggestion. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1577–1588. Association for Computational Linguistics (2011)
11. Liu, Z., Tu, C., Sun, M.: Tag dispatch model with social network regularization for microblog user tag suggestion. In: 24th International Conference on Computational Linguistics, p. 755. Citeseer (2012)
12. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to information retrieval, vol. 1. Cambridge University Press, Cambridge (2008)
13. Mei, Q., Cai, D., Zhang, D., Zhai, C.: Topic modeling with network regularization. In: Proceedings of WWW, pp. 101–110 (2008)
14. Peng, J., Zeng, D., Zhao, H., Wang, F.: Collaborative filtering in social tagging systems based on joint item-tag recommendations. In: Proceedings of CIKM, pp. 809–818. ACM (2010)
15. Rendle, S., Balby Marinho, L., Nanopoulos, A., Schmidt-Thieme, L.: Learning optimal ranking with tensor factorization for tag recommendation. In: Proceedings of KDD, pp. 727–736. ACM (2009)
16. Si, X., Sun, M.: Tag-LDA for scalable real-time tag recommendation. Journal of Computational Information Systems 6(1), 23–31 (2009)
17. Si, X., Liu, Z., Sun, M.: Modeling social annotations via latent reason identification (2010)
18. Symeonidis, P., Nanopoulos, A., Manolopoulos, Y.: Tag recommendations based on tensor dimensionality reduction. In: Proceedings of RecSys, pp. 43–50. ACM (2008)