From Raw Footprints to Personal Interests: Bridging the Semantic Gap via Trip Intention Aggregation

Long Guo [#1], Dongxiang Zhang [†2], Huayu Wu [*3], Bin Cui [§4], Kian-Lee Tan [‡5]

[#]School of EECS & Key Laboratory of High Confidence Software Technologies (MOE), Peking University
[§]School of Electronics and Computer Engineering (SECE), Peking University, SHENZHEN
[†]University of Electronic Science and Technology of China
[*]Institute for Infocomm Research, A*STAR [‡]National University of Singapore
{[1]guolong,[4]bin.cui}@pku.edu.cn, [2]zhangdo@uestc.edu.cn, [3]huwu@i2r.a-star.edu.sg, [5]tankl@comp.nus.edu.sg

***Abstract*—User-generated trajectories (UGT), such as GPS footprints from wearable devices or travel records from bus companies, capture rich information of human mobility and urban dynamics in the offline world. In this paper, our objective is to enrich these raw footprints and discover the users' personal interests by utilizing the semantic information contained in the spatial- and temporal-aware user-generated contents (STUGC) published in the online world.**

We design a novel probabilistic framework named CO^2 to connect the offline world with the online world in order to discover the users' interests directly from their raw footprints in UGT. In particular, we first propose a latent probabilistic generative model named STLDA to infer the intention attached with each trip, and then aggregate the extracted trip intentions to discover the users' personal interests. To tackle the inherent sparsity and noisiness problems of the tags in STUGC, STLDA considers the inner correlation between tags (i.e., semantic, spatial and temporal correlation) on the topic-level.

To evaluate the effectiveness of CO^2, we utilize a dataset containing three months of data with 5.3 billion bus records and a Twitter dataset with 1.5 million tweets published in 6 months in Singapore as a case study. Experimental results on these two real-world datasets show that CO^2 is effective in discovering user interests and improves the precision of the state-of-the-art method by 280%. In addition, we also conduct a questionnaire survey in Singapore to evaluate the effectiveness of CO^2. The results further validate the superiority of CO^2.

I. Introduction

With the rapid development of wireless communication technologies and GPS-enabled devices, there have been overwhelming amounts of user-generated trajectories (UGT) collected every day. These data, including public transport trajectories [1], travel trajectories [2] and life logging trajectories [3], contain rich information about human mobility patterns and urban dynamics, and have been applied to support various trajectory mining applications. However, since there is no textual information contained in UGT, these raw footprints cannot be used to support a wide class of location-aware personalized recommendation applications, in which the personal interests of the trajectory subjects need to be captured.

To bridge the semantic gap, in this paper, we attempt to design a framework to discover users' personal interests directly from their raw footprints in UGT effectively, which is very challenging since there is no textual information contained in UGT that can be utilized to facilitate the discovery process. Motivated by existing studies on trajectory semantic annotation, our solution is to first infer the trip intentions based on the rich semantic information contained in the spatial- and temporal-aware user-generated contents (denoted by STUGC) and then discover users' interests based on their intentions. The key assumption behind this solution is that the trip intentions are highly correlated with users' personal interests, because in reality a daily trip made by a user is often interest-driven such as the trip to a shopping mall or bar club. Therefore, if the trip intentions can be well inferred, the user's personal interests can be discovered with a high quality by aggregating the trip intentions.

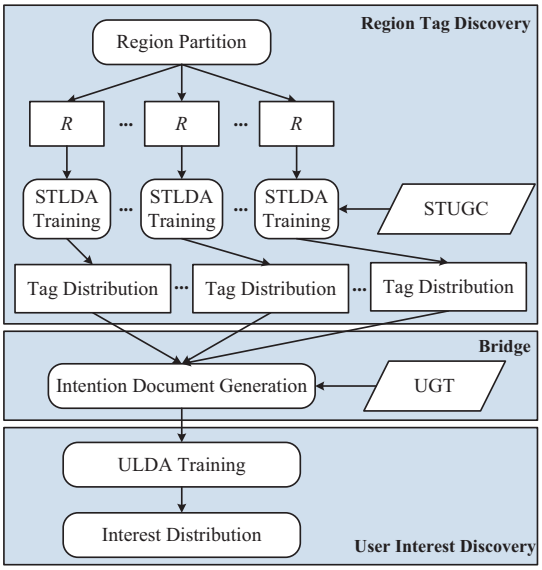

Fig. 1: Framework for Discovering the User Interests

However, there are several difficulties in inferring user interests from their raw footprints by using the trip intentions discovered from STUGC. First, the tags in STUGC suffer from severe sparsity problem. For instance, the tags that capture real trip intentions can be very sparse around the destination area. Second, the tags in STUGC are very noisy. For instance, a tweet about shopping may contain several tags irrelevant to the topic, which may overwhelm the real meaningful tags. Third, users' raw footprints are highly skewed. For example, a user can only visit a limited number of places. As a result, the trip intentions of this user belong to a sparse tag space, based on which the user's interests cannot be discovered thoroughly. Therefore, in order to infer the users' interests from their raw footprints with a high quality, we need to design an effective framework to overcome all these challenges.

2375-026X/17 $31.00 © 2017 IEEE
DOI 10.1109/ICDE.2017.55

To this end, we propose a novel probabilistic framework named CO^2 (i.e., connecting the offline world with the online world) to discover the user interests effectively. As shown in Fig. 1, there are three components in CO^2. The component on the top is used to discover the tag distributions for each region created by partitioning an urban city using a data-driven approach. To handle the sparsity and noisiness problems of the tags appearing in STUGC, we consider the inner correlation between tags (i.e., semantic, spatial and temporal correlation) on the topic-level instead of considering the tags independently. In particular, we propose a latent probabilistic generative model named STLDA to infer the distribution of tags based on the distribution of latent topics for each individual region. The advantage of inferring the distribution of tags on the topic-level are twofold. First, the sparsity problem of some meaningful tags can be well addressed due to its semantic relationship with the tags in the same topic. Second, the impact of the noisy tags in a message can be weakened significantly because in STLDA a message is characterized by topics instead of tags. The second component can be viewed as a bridge to connect the offline world with the online world which generates the trip intentions for the users based on their footprints and the inferred tag distributions of the regions. To better capture the trip intentions, a novel concept named "stay" is defined to take into account the spatial and temporal factors that influence users' visiting behavior. In particular, we consider the duration of a visit instead of the timestamp when inferring the trip intentions. The component at the bottom is used to discover the user interests in the form of weighted tags or topics by aggregating the generated trip intentions. To handle the skewed distribution of human movements, we propose a user-aware probabilistic generative model named ULDA to discover the interests for the users simultaneously.

II. Probabilistic Framework

In this section, we introduce the probabilistic framework according to the three components shown in Fig. 1.

A. Region Tag Discovery

The first component is used to infer the tag distributions for each region in the offline world with the help of STUGC published in the online world. There are many possible ways to partition an urban city into regions. We delete the presentation of region partition due to the limited space, which is available in [4]. In this section, we assume the city has already been partitioned into a set of regions and our goal is to infer the tag distributions for each region.

The key challenge in the first component is how to handle the sparsity and noisiness problems of the tags. Inspired by existing studies on probabilistic topic models, we propose a spatial- and temporal-aware LDA-based model named STLDA which adopts latent topics to characterize the tags for a region. The topic-based model clusters similar tags as a topic, and can put more weight on the topics with meaningful tags and less weight on the topics with noisy tags.

1) Model Structure: The graphical representation of STLDA is shown in Fig. 2. For each region r, we create a region corpus $\mathcal{D}_r$, which is a collection of spatial- and temporal-aware social messages, such as geo-tagged tweets and check-ins, located within r. Each message d in $\mathcal{D}_r$ comprises a triple

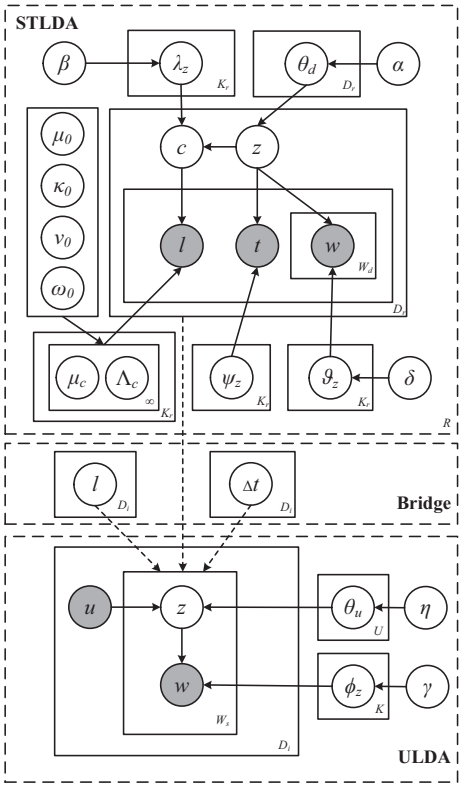

Fig. 2: The Graphical Representation

$(\mathcal{W}_d, t, l)$, where $\mathcal{W}_d$ denotes the set of semantic tags, t denotes the publish time and l denotes the publish location. We use $\mathcal{W}$ to denote the set of the tags in $\mathcal{D}_r$, and define $|\mathcal{D}_r| = D_r$ and $|\mathcal{W}_d| = W_d$. Given $\mathcal{D}_r$ as input, STLDA discovers the topics for region r by considering the semantic tags, timestamps and geographical locations in the messages jointly. In the following, we describe each component in STLDA.

Document-topic distribution $\boldsymbol{\theta_d}$. Each message d in $\mathcal{D}_r$ is represented as a random mixture over latent topics $\boldsymbol{z}$ and characterized as a multinomial distribution over $\boldsymbol{z}$, i.e., $\theta_d = \{\theta_{d,z} : z \in \boldsymbol{z}\}$ where $\theta_{d,z}$ denotes the probability of message d choosing topic z and $|\boldsymbol{z}| = K_r$. In this way, we can identify the real topic of d and reduce the negative impact coming from the irrelevant tags in d.

Topic distribution. To better extract the representative tags of a region, it is important to discover the hidden topics with a good quality. In STLDA, a topic z is not only responsible for generating the tags in a message, but also correlated with its time and location. Next we discuss each type of correlation.

I. Topic-tag distribution $\boldsymbol{\vartheta_z}$. A topic z is defined as a multinomial distribution over $\mathcal{W}$, i.e., $\vartheta_z = \{\vartheta_{z,w} : w \in \mathcal{W}\}$ where $\vartheta_{z,w}$ denotes the probability of topic z generating tag w. Since a social message is very short and usually belongs to only one topic, we follow this assumption in STLDA and assign all the tags in a tweet to the same topic.

II. Topic-time distribution $\boldsymbol{\vartheta_z}$. The topic of a message has a high correlation with the publish time of the message. For example, tweets about restaurants are usually published at lunch or dinner time, while tweets about bars are more

frequently published at midnight. Therefore, we assume that the topic has a distribution over time as well. Unlike the topic-tag distribution, we do not simply use the discretization method due to the continuity of the time. Instead, we normalize the time during a day to a range from 0 to 1 and adopt the Beta distribution which can simulate various shapes to model the distribution of the topic over time, i.e., $t \sim Beta(\psi_{z,1}, \psi_{z,2})$. Compared with the discretization method, the continuous representation of the time can capture the temporal distribution of the tags in a more flexible and precise way.

III. Topic-location distribution $\boldsymbol{\lambda_z}$. Similarly, tags used in a geo-tagged message depend heavily on the publish location of the message. For example, tweets published within a bar street are more likely about wine, while those published in a shopping mall are more likely about shopping. As a result, the topics within a region often follow a similar distribution as the underlying POIs. Like the topic-time distribution, we also want to capture the continuity of the locations instead of using the discretization method, which is typically captured by the Gaussian distribution. However, one topic may be distributed at multiple areas. For example, there may be several shopping malls within a region. Therefore, there should be several centroids corresponding to the topic about shopping within a region. Based on this observation, we use the Gaussian mixture model (GMM) to represent the distribution of a topic. In particular, a topic z is defined as a multinomial distribution over clusters $\mathcal{C}_z$, i.e., $\lambda_z = \{\lambda_{z,c} : c \in \mathcal{C}_z\}$ where $\lambda_{z,c}$ denotes the probability of topic z belonging to cluster c and $|\mathcal{C}_z| = C_z$. Within each cluster c, the location distribution is captured via the Gaussian distribution, i.e., $l \sim \mathcal{N}(\mu_c, \Lambda_c^{-1})$. Since different topics may have different number of clusters, it is important to find the appropriate number of clusters for each topic. To achieve this goal, we employ Chinese Restaurant Process (CRP) to generate the clusters. CRP is a stochastic process where a customer can either sit at an occupied table i with probability of $\frac{n_i}{n+\beta-1}$ or a new table with probability of $\frac{\beta}{n+\beta-1}$ after the first customer randomly picks a table, where n_i is the number of customers at table i. As a Bayesian non-parametric approach, CRP can automatically estimate how many clusters are needed to model the data.

2) Generative Process: As presented in Fig. 2, STLDA assumes the following generative process for each message d in a region corpus $\mathcal{D}_r$: [1]:

1) Choose $\theta_d \sim Dir(\alpha)$.
2) Choose a topic $z \sim Multi(\theta_d)$
3) With the chosen topic z,
    a) choose $\vartheta_z \sim Dir(\delta)$,
    b) choose tag $w \sim Multi(\vartheta_z)$ for each $w \in \mathcal{W}_d$,
    c) choose time $t \sim Beta(\psi_{z,1}, \psi_{z,2})$,
    d) choose cluster c based on the CRP,
    e) with the chosen cluster c,
        i) choose $\{\mu_c, \Lambda_c\} \sim \mathcal{NW}(\mu_0, \kappa_0, \upsilon_0, \omega_0)$,
        ii) choose location $l \sim \mathcal{N}(\mu_c, \Lambda_c^{-1})$.

where $\mathcal{NW}(\cdot)$ denotes the Normal Wishart distribution.

[1]To avoid overfitting, we place a Dirichlet prior over each multinomial distribution.

3) Model Inference: For each region r, our goal is to learn the model parameters to maximize the marginal log-likelihood of the observed random variables $\boldsymbol{\mathcal{W}_d}, \boldsymbol{t}$ and $\boldsymbol{l}$. However, inference cannot be done exactly in STLDA. Thus, we employ the collapsed Gibbs sampling to estimate the unknown parameters $\{\boldsymbol{\theta_d}, \boldsymbol{\lambda_z}, \boldsymbol{\vartheta_z}, \boldsymbol{\psi_z}, \boldsymbol{\mu_c}, \boldsymbol{\Lambda}_c\}$. The idea of the collapsed Gibbs sampling is to iteratively sample the latent topics $\boldsymbol{z}$ and clusters $\boldsymbol{c}$ from a Markov Chain. Since we adopt conjugate prior (Dirichlet) for the multinomial distributions, we can easily integrate out $\boldsymbol{\theta_d}$, $\boldsymbol{\lambda_z}$ and $\boldsymbol{\vartheta_z}$ and need not sample them at all. Finally, the unknown parameters can be estimated from the samples. Due to the limited space, the detailed model inference procedure is deleted, which is available in [4].

B. Trip Intentions Generation

The second component is used to generate the trip intentions for the users based on their historical trips, which serves as an important bridging component between the offline and online worlds. As shown in Fig. 1, this component can be viewed as a bridge to connect the offline world where people make trips and the online world where people publish messages. A key observation in this paper is that the trip intentions are highly correlated with users' interests. For example, if a user often visits a place for its bars, we can say that this user likes bars with a high probability. If the trip intentions can be well inferred, the user's personal interests can be discovered with a high quality by aggregating the trip intentions. Therefore, our problem becomes how to precisely infer the trip intentions which is represented as a set of weighed tags semantically.

With the help of STLDA, we can discover the tag distribution of a region r. If a user visits r, a straightforward approach is to use the associated tag distribution to infer the trip intention. To improve the accuracy, we take into account two additional factors that influence users' visiting behavior. The first one is the temporal influence. For example, if a user visits a shopping mall at noon, she probably goes there for lunch. However, if she stays there during $12:00 \sim 16:00$, she is more likely to go there not only for lunch but also for shopping. Therefore, we consider the duration of a visit instead of the timestamp when inferring intentions, which can be achieved because STLDA uses the Beta distribution to capture the continuity of the time. The second one is the spatial influence. For example, a user is more likely to choose a location near the destination of her trip. Therefore, the tags distributed near the destination should be given higher priority, which can be achieved by the Gaussian mixture model adopted in STLDA. Next, we define a concept named "stay" to take into account this two factors.

Definition 1 (Stay): Let $\Delta t = (t_a, t_l)$. A stay $S_{u,r}$ is a quadruple $(u, r, \Delta t, l)$, which represents a user u arrives at a location l in a region r at time $\Delta t.t_a$ and leaves at time $\Delta t.t_l$.

The temporal influence and spatial influence of a stay can be integrated into the intention discovery process in a unified way. Given a stay $S_{u,r}$, we first choose the region r where $S_{u,r}$ belongs to. Then we can infer the trip intentions behind $S_{u,r}$, which is represented by a distribution over the tags in r. The probability for each tag can be derived using Eq. 1 based on the model parameters inferred for r in the first component.

With the help of these model parameters, we can infer the probability of a tag with any time duration and location.

$$
\begin{aligned}
P(w|\Delta t, l, \cdot) &= \frac{P(w, \Delta t, l|\cdot)}{\sum_{w' \in \mathcal{W}} P(w', \Delta t, l|\cdot)} \propto P(w, \Delta t, l|\cdot) \\
&= \sum_{z=1}^{K_r} P(w|z, \cdot) P(\Delta t|z, \cdot) \sum_{c=1}^{C_z} P(l|c) P(c|z, \cdot) \sum_{d=1}^{D_r} P(z|d, \cdot) P(d|\cdot) \\
&\propto \sum_{z=1}^{K_r} \sum_{c=1}^{C_z} \sum_{d=1}^{D_r} P(w|z, \cdot) \int_{t_a}^{t_l} P(t|z, \cdot) dt \; P(l|c) P(c|z, \cdot) P(z|d, \cdot) \\
&= \sum_{z=1}^{K_r} \sum_{c=1}^{C_z} \sum_{d=1}^{D_r} \phi_{z,w} \cdot \frac{B(t_l; \psi_{z,1}, \psi_{z,2}) - B(t_a; \psi_{z,1}, \psi_{z,2})}{B(\psi_{z,1}, \psi_{z,2})} \\
&\cdot \lambda_{z,c} \cdot \frac{1}{2\pi \sqrt{|\Lambda_c^{-1}|}} exp(\frac{-(l - \mu_c) \Lambda_c (l - \mu_c)}{2}) \cdot \theta_{d,z}
\end{aligned}
\tag{1}
$$

where $B(t; \psi_{z,1}, \psi_{z,2})$ denotes the incomplete Beta function.

Based on the stays, we can create an intention corpus $\mathcal{D}_i$ containing the trip intentions of all users, where each intention document s is created by one stay according to Eq. 1 and composed of a user ID u and a set of weighted tags $\mathcal{W}_s$. We define the number of users as U, $|\mathcal{D}_i| = D_i$ and $|\mathcal{W}_s| = W_s$.

C. Trip Intention Aggregation

The third component is used to discover the users' interests represented by a set of weighed tags or topics by aggregating the trip intentions in $\mathcal{D}_i$. To discover the interests of a user u, a naive method is to consider the tag or topic distributions within the intention documents of u. However, it is challenging to discover the interests of a user by only considering her intention documents, because in reality a user often visits a limited number of places, resulting in a sparse user-tag or user-topic matrix. To alleviate the sparsity problem, we borrow the idea of traditional item-based collaborative filtering method. If a user is interested in one tag, she may be also interested in other tags similar with that tag. Based on this idea, we discover the interests for the users by considering their intention documents simultaneously and use the latent topics to characterize the similarity between tags. In this way, given a user, we can use some tags appearing in the intention documents of other users to annotate this user. As a result, the sparsity problem can be alleviated.

In particular, we build a user-aware LDA-based model named ULDA as presented in Fig. 2 to model the interests for the users simultaneously. The difference between ULDA and traditional LDA is that we define the user instead of the document to be a random mixture over latent topics. As a result, ULDA assumes the following generative process for each document s in $\mathcal{D}_i$:

1) Choose $\theta_u \sim Dir(\eta)$.
2) For each tag $w \in \mathcal{W}_s$:
   a) Choose a topic $z \sim Multi(\theta_u)$.
   b) Choose $\phi_z \sim Dir(\gamma)$.
   c) Choose a tag $w \sim Multi(\phi_z)$.

where θ_u represents the distribution of the user u over topics.

We can also use the collapsed Gibbs sampling method to infer the unknown parameters $\boldsymbol{\theta_u}$ and $\boldsymbol{\phi_z}$. However, the other difference of ULDA compared with LDA is that the tags in each intention document $s \in \mathcal{D}_i$ are weighted. In LDA, all the tags have equal importance and thus the tag frequency is considered when sampling the topics. While in ULDA, the tags have different weight used to indicate different importance. Therefore, we propose a weight-based way to calculate the conditional probability $P(z_i = z|\boldsymbol{z}_{-i}, \cdot)$ as follows:

$$
P(z_i = z|\boldsymbol{z}_{-i}, \cdot) \propto (N_{u,z,-i} + \eta) \cdot \frac{N_{z,w,-i} + \gamma}{N_{z,-i} + W_s \gamma} \tag{2}
$$

where $N_{u,z,-i}$ is the sum of the weights of the tags in the documents belonging to u assigned to topic z, $N_{z,w,-i}$ is the sum of the weights of the instances of tag w assigned to topic z and $N_{z,-i}$ is the sum of the weights of all the tags assigned to topic z. In other tags, when a tag is assigned to a topic, it contributes the topic with its importance. After a sufficient number of iterations, we can estimate $\boldsymbol{\theta_u}$ and $\boldsymbol{\phi_z}$ as follows:

$$
\begin{aligned}
\theta_u &= P(z|u) = \frac{N_{u,z} + \eta}{\sum_{z' \in \boldsymbol{z}} (N_{u,z'} + \eta)} \\
\phi_z &= P(w|z) = \frac{N_{z,w} + \gamma}{\sum_{w' \in \mathcal{W}_s} (N_{z,w'} + \gamma)}
\end{aligned}
\tag{3}
$$

Now we have a distribution of user interests over topics, i.e., θ_u. We can also infer the distribution of user interests over tags as:

$$
P(w|u) = \sum_{z=1}^{K} P(w|z) p(z|u) = \sum_{z=1}^{K} \phi_z \cdot \theta_u \tag{4}
$$

To test the effectiveness of CO^2, extensive experiments have been done on two large-scale and real-word datasets. For more information about the experiment settings and experiment results, please refer to [4].

III. Conclusion

In this paper, we attempt to design a probabilistic framework named CO^2 to connect the offline world with the online world in order to discover the interests for the users based on their raw footprints in UGT effectively. To evaluate the effectiveness of CO^2, we use two large-scale and real-world datasets as a case study. Experimental results show that CO^2 is effective in discovering user interests.

IV. Acknowledgement

This research is supported by the National Natural Science Foundation of China under Grant No. 61572039 and Grant No. 61602087, 973 program under No. 2014CB340405, Shenzhen Gov Research Project JCYJ20151014093505032, and NRF-NSFC Joint Grant NRF2016NRF-NSFC001-113.

References

[1] N. Yuan, Y. Wang, F. Zhang, X. Xie, and G. Sun, "Reconstructing individual mobility from smart card transactions: A space alignment approach," in *ICDM*, 2013.

[2] S. Counts and M. Smith, "Where were we: Communities for sharing space-time trails," in *GIS*, 2007.

[3] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma, "Mining interesting locations and travel sequences from gps trajectories," in *WWW*, 2009.

[4] L. Guo, D. Zhang, H. Wu, b. Cui, and K.-L. Tan, "From raw footprints to personal interests: Bridging the semantic gap via trip intention aggregation," 2017. [Online]. Available: http://net.pku.edu.cn/ daim/guolong/files/ICDE.pdf