Modeling User Intrinsic Characteristic on Social Media for Identity Linkage

Xianqi Yu
Shandong University
Jinan, China
xqyu1993@126.com

Yuqing Sun
Shandong University
Jinan, China
sun_yuqing@sdu.edu.cn

Elisa Bertino
Purdue University
West Lafayette, USA
bertino@cs.purdue.edu

Xin Li
Shandong University
Jinan, China
lx@sdu.edu.cn

ABSTRACT

Most users on social media have intrinsic characteristics, such as interests and political views, that can be exploited to identify and track them. It raises privacy and identity issues in online communities. In this paper we investigate the problem of user identity linkage on two behavior datasets collected from different experiments. Specifically, we focus on user linkage based on users' interaction behaviors with respect to content topics. We propose an embedding method to model a topic as a vector in a latent space so as to interpret its deep semantics. Then a user is modeled as a vector based on his or her interactions with topics. The embedding representations of topics are learned by optimizing the joint-objective: the compatibility between topics with similar semantics, the discriminative abilities of topics to distinguish identities, and the consistency of the same user's characteristics from two datasets. The effectiveness of our method is verified on real-life datasets and the results show that it outperforms related methods.

ACM Classification Keywords

J.4 Social and Behavioral sciences

Author Keywords

Social Media; Privacy Issue; Identity Linkage; Intrinsic Characteristic; Embedding Method.

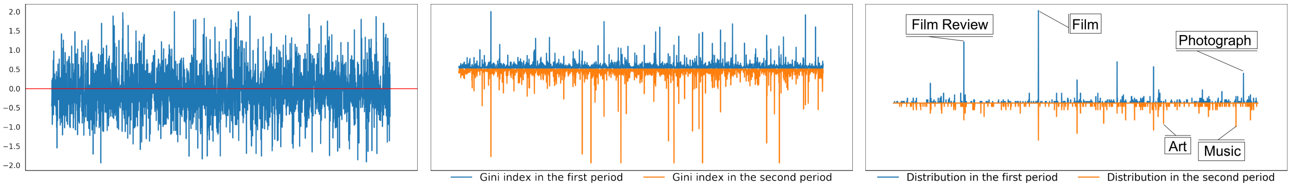
INTRODUCTION

The User Identity Linkage (UIL) problem refers to the problem of recognizing that two user identities from two different data sources actually refer to the same individual in real life [25]. The problem has recently attracted an increasing amount of attention from both academia and industry. It is a special concern of social communities because of privacy issues. It is also of critical importance for service providers to have deeper understanding of their customers from multiple perspectives so as to profile users on social media for better promotions or services.

In this paper, we focus on a setting in which we are given user behaviors collected during two time periods on the same social media platform. It is a common setting in UIL approaches that link identities via behavior data [8, 9, 27, 29]. It is assumed that, in the first period, user identities are known and in the second period their identities are anonymized [22]. For example, user identities might be changed to anonymous after applying some identity management system for protecting against identity linkage attack [10]. A user's web behavior may refer to browsing a piece of news on CNN, rating a film on MovieLens, or answering a question on Quora. Generally, such behaviors are relevant to some topics. For example, a user answered the question *How to evaluate Donald J. Trump be elected the 45th president of the United States* on Quora, and topics of this behavior can be tags *Political* and *President Election* marked on the question. We regard these interactive activities between users and related topics as *interactions* and denote a user behavior as a set of topics here afterwards. Our goal is to answer whether users can be identified only by their interactions with topics. Since the above settings are widely available in practice, this work would provide meaningful results for many related applications.

The identification of users based on their topic interactions requires addressing two challenges. One is the dynamic evolution of popular topics. We collected some statistics on topic frequency among all user behaviors in two time periods and obtained two probability distributions over topics. Figure 1(a) shows the difference between two probability distributions. The X-axis represents topics and the Y-axis represents the changing ratio of topic proportion, which is calculated as their difference divided by the average of topic proportion in two periods. The red line can be seen as a reference for the case in which the popularity of topic does not change between two pe-

GROUP 2018, January 7–10, 2018, Sanibel Island, FL, USA
Copyright © 2018 Association of Computing Machinery.
ACM ISBN 978-1-4503-5562-9/18/01 ...$15.00.
http://dx.doi.org/10.1145/3148330.3148340



(a) Changing ratio of topic proportion over two periods

(b) Gini index of topics in two periods

(c) Probability distributions over topics of a selected user in two periods

Figure 1. The data are collected from Zhihu website which consist of users' behavior data in a whole year period. Figure(a) shows the changing ratio of topic proportion in two periods. Figure(b) shows the Gini index of topics in two periods. Figure(c) shows, for a randomly selected user, two probability distributions over his or her interested topics in two periods.

riods. The results show that most topics have obvious changes, either in an increasing trend or in a decreasing trend. Thus it is clear that popular topics vary a lot in different periods. As our problem is to distinguish user identities based on their behaviors, we also evaluate the discriminative abilities of topics. Considering the fact that many users pay almost the same attentions on some specific topics, it is difficult to distinguish them against these topics. However, if a topic is of concern to only a few users, this topic would be helpful to identify them. So we adopt Gini index to evaluate the discriminative abilities of topics, which is calculated as $1-\sum_{i=1}^{m} p_i^2$, where p_i denotes the probability of the i^{th} topic. It is often used as a measure of the impurity of data [12]. Figure 1(b) shows the Gini indexes of topics in two periods, where the X-axis represents topic and the Y-axis represents the Gini index. The higher Gini index, the higher the discriminative ability of the topic. In the figure, the upper half is the Gini index in the first period and the lower half is for the second period. The great difference between the two periods indicates that for each topic, the discriminative ability on user identities has changed over time.

Another observation is the change of topics of interest for each user. As shown in Figure 1(c), for a randomly selected user, the probability distributions over topics are quite different in two different periods, as shown by the difference between the blue and orange line in the figure. For example, the peak topics that this user has interacted with in the first period are *film review*, *film*, *photograph* etc., while in the second period, they changed to *art*, *music*, etc, which are marked in the figure. It is easy to understand that the statistics over topics in two periods are quite different and thus two identities cannot be linked by the similarity on the statistics over topics of interests.

To address the above challenges, this paper introduces the concept of user intrinsic characteristic so as to identify the inner motivations implied in user behaviors. Consider the example in Figure 1(c). Note that, although the topics of interest for the user seem different, it does not mean that the user's interests have changed. From the point of view of an art lover, the characteristics of this user remain the same in these behaviors. Thus our approach is to investigate the implicit semantics of topics by embedding each topic as a vector in a latent space and model users' intrinsic characteristics based on their interactions with topics. A topic representation is denoted by a d-dimensional vector, which can be learned from the training data. A user vector is the statistics on topics and then is mapped to the same latent space as topics. In order to learn the embedding representations of both topics and user characteristics, we apply a joint-objective optimization. The first optimization objective is to maximize the compatibility between topics with similar semantics and their discriminative ability with respect to identity recognition. We define the semantic relevance between a pair of topics as their compatibility. Topics co-occurring in a behavior share a higher compatibility score, and thus their representations should be close in the latent space. The discriminating abilities of topics reflect the differences of user interactions on topics. We take the correlation between topics as a regularization in vector compatibility learning. The second optimization objective is to maximize the consistency of two characteristics in different time periods for each seed user, which reflects the fact that the intrinsic characteristics of an individual often remain stable over time. Then we adopt the learned embedding vectors to solve the task of identity linkage. We conduct experiments on two real-life datasets, and the results show that our method outperform related methods.

The rest of the paper is organized as follows. Section 2 discusses related works. Sections 3 and 4 present the formal definition of the UIL problem and the model of user intrinsic characteristics, respectively. We discuss in Section 5 the learning process of embedding representations. Section 6 presents the experimental results and Section 7 concludes the paper.

RELATED WORKS

Privacy & Identity Issues in Online Communities

The privacy and identity issues in online communities have attracted an increasingly amount of attention after the emergence of de-anonymization techniques, which match users in an anonymous dataset to real individuals. In recent years, organizations release more and more datasets in online communities for problem solving, such as business promotion or epidemic diseases prediction. These datasets contain many sensitive individual information, such as health histories or transactions. Although these datasets are anonymized by some techniques such as k-anonymization [6], recent research has shown that an adversary can use auxiliary information to de-anonymize users' records from the correlated and publicly available datasets [23, 24]. Narayanan et al. first introduced this problem [23], and used film reviews in IMDB as auxiliary information to successfully re-identify a number of specific users in anonymous Netflix data via comparing user activities in the two datasets. They also proposed a de-anonymization algorithm on user relation networks, which is able to effectively re-identify users in the anonymized graphs with only a few auxiliary information [24]. All these methods are based on the the presence of overlapping information between the anonymous and auxiliary datasets. Unlike such work, this paper focuses on the setting in which users' data are collected from two non-overlapping time periods. A different way of thinking to model users' intrinsic characteristics is presented to conduct de-anonymization.

Behavior Based Identity Linkage

Behavior based identity linkage is becoming a promising problem in the field of social computing in recent years. Many papers report results of analyses based on statistic methods to address this problem. Zang et al. performed a study on a nationwide call-data record dataset, and demonstrated that the most frequently visited locations can act as quasi-identifiers to re-identify users [29]. Gambs et al. introduced a Markov model to analyze the temporal evolution of the mobility patterns of the users [8, 9]. These data are all relevant to user mobile intelligent devices and reflect users' physical movements or real contacts. Comparatively, in our problem, user behaviors on social media are more noisy and random, which make the above methods inapplicable to our setting. Unnikrishnan et al. proposed a statistical method for matching user identity based on browsing history [22, 27]. They preprocess item data as categorical types and model user behavior as the statistics on these categories, where each user is formalized as a distinguished probability distribution pattern. The assumption behind such an approach is that each dimension of a random vector is independent from the others and each behavior follows an independent and identical distribution. However, in practice, such assumption does not always hold. For example, suppose that on a news there are the following topics: *Presidential Election*, *Trump* and *Political*. These topics are regarded as categorical data in the probability distribution but they are semantically related. Moreover, the behavior of reading a news about *Presidential Election* is probably followed by the behavior of reading a news about *Trump's Speech*, which is not an independent and identical distributed trial. Understanding semantics of people's behaviors on social media sites is a complex task, requiring a series of systematic studies. Bakhshi et al. examine the relationship between social signals and the emotional valence of users' reviews on the online recommendation community Yelp [2]. Some methods in collaborative recommendation systems model users' preference on the Web as a latent semantic vector by matrix factorization [4, 15]. Similar to this idea of finding latent factor, we model users' intrinsic characteristics in a latent vector space. Compared to these works, we learn the latent representations of users by a quite different objective function.

User Linkage Across Social Media Platforms

A lot of research has focused on the problem user linkage across social networks, which are highly related to our work. In social networks, information about user attributes and user relation network can be used to link user identity across different social platforms. Some researchers have demonstrated that it is possible to recognize user identities by the structure of their social networks. Korula and Lattanzi introduced a many-to-many mapping algorithm based on the degrees of unmapped users and the number of common neighbors with the help of anchor users [16]. Bartunov et al. proposed an approach based on the conditional random fields called Joint Link-Attribute (JLA) [3], which considered both profile attributes and network properties. Liu et at. proposed a heterogeneous behavior modeling method [18]. They combined user attributes, topic distribution (obtained by LDA [5], which is a generative probabilistic model for collections of discrete data such as text corpora) and graph topology, and other information, and learn the mapping function by a multi-objective optimization to match user accounts from different social networks. Although these approaches show that jointly using user attributes and network structure can lead to better performance, such information is often unavailable in many online communities. So they are not appropriate for solving our problem. Amitay et al. studied the problem of author detection over a collection of blog pages originating from different sources and written to serve different online functions [1]. They proposed a compress based method to solve the problem. Different from focusing on the User Generated Content(UGC), we want to show how much the topics in users' web behaviors can reveal their identities. We propose an embedding method which focuses on interpreting the semantics of user behaviors. Then we are able to model users' intrinsic characteristics and identify users.

Table 1. Notations in this paper

SYMBOL	DESCRIPTION
u	user
U	set of users
T	set of topics
t_i	the i^{th} topic in T
B_u	behavior sequence of user u
$\mathbf{b}_i$	the i^{th} behavior in B_u
$\mathbb{B}$	set of behavior sequences of all users
$\mathbf{d}_u$	probability distribution over topics for user u
$\mathbf{v}_i$	embedding representation of topic t_i
$\mathbf{V}$	embedding matrix of topics in T
$\mathbf{p}_u$	intrinsic characteristic vector for user u
e	event of topic co-occurrence
c	normalization parameter for soft-max probability function
θ	parameters to be learned, including $\mathbf{V}$ and c
λ	weight parameter of regularization term
γ	preference parameter in joint objective optimization function

PROBLEM DEFINITION

Let U denote the user set in a given setting. For any user $u \in U$, his or her behaviors on social media are given as a sequence of topic interactions. Let $T = \{t_1, t_2, ..., t_{|T|}\}$ represent the set of all topics on a platform. The behavior sequence of u is denoted as $B_u = [\mathbf{b}_1, ..., \mathbf{b}_{|B_u|}]$, where each behavior $\mathbf{b}_i$ is a vector of size $|T|$, $\mathbf{b}_i \in \{0,1\}^{|T|}$. For each behavior $\mathbf{b}_i$, $\mathbf{b}_i(k) = 1$

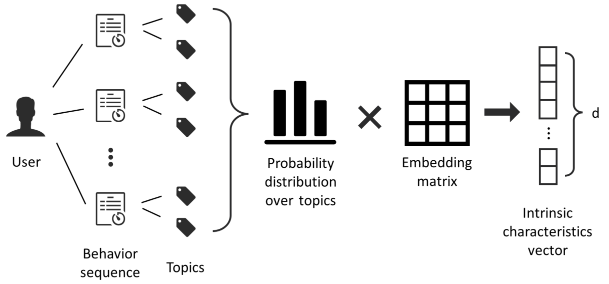

User
Behavior sequence
Topics
Probability distribution over topics
Embedding matrix
Intrinsic characteristics vector
d

Figure 2. Modeling user intrinsic characteristic based on topic embedding

indicates that $\mathbf{b}_i$ interacts with topic t_k; And $\mathbf{b}_i(k) = 0$ otherwise. Let $\mathbb{B}_1 = \{B_1, B_2, ..., B_{|\mathbb{B}_1|}\}$ and $\mathbb{B}_2 = \{B_1, B_2, ..., B_{|\mathbb{B}_2|}\}$ denote two sets of behavior sequences collected from two separate time periods. We summarize the notations used in this paper in Table 1. The UIL problem is defined as follows.

Definition 1. **User Identity Linkage (UIL)**: Given a set of users U, and their behavior sequences from two time periods: the identity-labeled $\mathbb{B}_1$ and the anonymized $\mathbb{B}_2$, the UIL problem is to label behavior sequences in $\mathbb{B}_2$ with user identities in $\mathbb{B}_1$.

INTRINSIC CHARACTERISTIC MODELING

From the above discussion, we can see the challenges to UIL are topic popularity evolution and variations of the similar topics. To solve these challenges, we propose a topic embedding based user intrinsic characteristics model as illustrated in Figure 2. The model includes two parts: learning topic representation in latent space according to a joint-objective optimization and modeling user intrinsic characteristic against behavior related topics. Based on the intrinsic characteristics, we then verify user identity mapping relationships based on user vectors in the latent space.

To model user intrinsic characteristics, we first learn user behaviors by statistics over topics. For a user's behavior sequence B_u, let $\mathbf{d}_u \in R^{|T|}$ denote the probability distribution over topics, where the k^{th} element of $\mathbf{d}_u$ is given by:

$$\mathbf{d}_u(k) = \frac{\sum_{i=1}^{|B_u|} \mathbf{b}_i(k)}{\sum_{i=1}^{|B_u|} \sum_{j=1}^{|T|} \mathbf{b}_i(j)} \quad ,k = 1, 2, ..., |T|. \tag{1}$$

To interpret the semantics of topics, we embed them into a latent space. Each topic is represented as a d-dimensional vector representing some intrinsic characteristics. Let matrix $\mathbf{V} \in R^{|T| \times d}$ denote the embedding representations of topics,

$$\mathbf{V} = \begin{bmatrix} \mathbf{v}_1^\top \\ \mathbf{v}_2^\top \\ \vdots \\ \mathbf{v}_{|T|}^\top \end{bmatrix} \tag{2}$$

where $\mathbf{v}_i$ is the embedding representation of topic t_i. Then a user's intrinsic characteristic is modeled as a linear transformation of the topic distribution $\mathbf{d}_u$, namely $\mathbf{p}_u = \mathbf{V} \cdot \mathbf{d}_u$. Here the topic embedding matrix $\mathbf{V}$ is called the transformation matrix. Modeling user intrinsic characteristic has two benefits. From the perspective of a single user, it helps in finding the common semantics in the dynamics of the topics of interest to the user and so to keep the consistency of one's traces in different time periods. From the global perspective, it helps interpreting the semantics of a newly emerged topic. Besides, since many topics are created by users, they might be noisy and sparse. The embedding method can reduce the dimension of topic space in user behaviors.

Based on the topic vectors, the UIL problem can be solved by three steps: 1)To model the intrinsic characteristics of each user behavior sequence in both $\mathbb{B}_1$ and $\mathbb{B}_2$; 2) To quantify the similarity between two user vectors $\mathbf{p}_u$ from $\mathbb{B}_1$ and $\mathbf{p}'_u$ from $\mathbb{B}_2$; 3) For a target anonymized user $\mathbf{p}'_u$, to use the nearest neighbor method to find the top k similar users in $\mathbb{B}_1$. There are many candidate distance functions, such as Euclidean distance and Cosine distance. We would discuss in details which is appropriate for the UIL problem in the experiments section .

EMBEDDING LEARNING

In this section, we first discuss the joint-objective of the topic embedding learning process and then present the learning algorithm.

Joint-Objective

The embedding representations are learned by jointly optimizing two objectives. The first objective is to maximize the compatibility between topics. This is motivated by the fact that topics associated with the same content are often related. For example, on the Q&A website Quora, a user answered the question *How to learn deep learning*. The tags marked by users on the question are regarded as topics, such as *Machine Learning* and *Deep Learning*. Although they are different words, they are actually highly related with respect to semantics. That is, topics co-occurring in a behavior always have high compatibility. Consequently, their embedding representations should be close in the latent space. So we introduce the compatibility score between a pair of topics as their semantic relevance.

The co-occurrence of topics t_i and t_j is defined as an event e_{ij}. The compatibility score of e_{ij} is given by:

$$S_\theta(e_{ij}) = \mathbf{v}_i \cdot \mathbf{v}_j \tag{3}$$

where $\theta = \{\mathbf{V}\}$ denotes the set of model parameters.

When we consider the topic compatibility with respect to semantics, at the same time, we also take into account the discriminating ability on identity linkage, that is, how much a pair of topics contribute in distinguishing identities. We thus introduce the correlation coefficient between two topics as an adjustment parameter, which is learned from the statistics on these topics against all user behaviors. There are many correlation function candidates. For example, the Pearson correlation coefficient (PCC)[1] can be chosen as a measure, which is the linear correlation between two variables and ranges from -1 to 1, where value 1 indicates they are totally positive linearly correlated, and value -1 indicates they are totally negative linearly correlated. Let PCC_{ij} denote the PCC of topic t_i and t_j.

[1] https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

We refine the compatibility score between topics t_i and t_j as:

$$S_\theta(e_{ij}) = \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{\sigma\Big(\mathrm{PCC}_{ij}\Big)} \tag{4}$$

where $\sigma(x) = \frac{1}{1+\exp(-x)}$. The introduction of the sigmoid function $\sigma(x)$ is to prevent the denominator from being zero.

Here is an example to illustrate why PPC is helpful for the discriminative ability of a pair of topics. Consider the pair of topics *Dota* and *LOL*, two popular computer games of the same type. They are occasionally mentioned together for comparison and discussion purposes. But in a more general case, they appear independently. Based on general knowledge about the game field, a *Dota* game player seldom plays game *LOL*, and vice versa. If we learn their representations only by their co-occurrences, they would be very close to each other and it would be difficult to distinguish two types of players. Since these two topics share a low PCC, by using the correlation parameter, their compatibility score can also reach a high value without their representations being too close.

Let $\mathbf{E}$ be the set of events, namely all topic pairs, $|\mathbf{E}| = \frac{|T|(|T|-1)}{2}$. We adopt the soft-max function to model the occurrence probability of such an event.

$$P_\theta(e) = \frac{\exp(S_\theta(e))}{\sum_{x\in\mathbf{E}}\exp(S_\theta(x))} \tag{5}$$

Let E_p denote the event dataset of all pair-wise topic co-occurrences extracted from the training data. The loss function of topic compatibility is defined as follows:

$$J_T(\theta) = -\sum_{e\in E_p}\log P_\theta(e) \tag{6}$$

The second objective is to maximize the consistency of the intrinsic characteristics of the same user. Recall the notions of user behavior sequences in two periods, $B_u \in \mathbb{B}_1$ and $B'_u \in \mathbb{B}_2$, respectively. For any two sequences B_u and B'_u belonging to the same user u, the corresponding latent vectors are $\mathbf{p}_u$ and $\mathbf{p}'_u$. Let $\mathrm{Dist}(\mathbf{p}_u, \mathbf{p}'_u) = -\mathbf{p}_u \cdot \mathbf{p}'_u$ evaluates the distance between them. Given a set of seed users labeled in both periods, denoted as $U_{\mathrm{seed}} \subset U$, and B_u and B'_u from two periods belonging to the same user $u \in U_{\mathrm{seed}}$, our goal is to maximize the consistency of the same user and the difference between different users. The objective function for minimization is defined as follows:

$$J_C(\theta) = \sum_{u\in U_{\mathrm{seed}}, v\in U, u\neq v}\Big(\mathrm{Dist}(\mathbf{p}_u,\mathbf{p}'_u) - \mathrm{Dist}(\mathbf{p}_u,\mathbf{p}'_v)\Big) \tag{7}$$

We transform the function into the form of hinge loss and add a regularization term:

$$\begin{aligned} J_C(\theta) = \sum_{u\in U_{\mathrm{seed}}, v\in U, u\neq v}\Big(&\max(0,\ \mathrm{Dist}(\mathbf{p}_u,\mathbf{p}'_u) - \\ &\mathrm{Dist}(\mathbf{p}_u,\mathbf{p}'_v) + \varepsilon)\Big) + \lambda||\mathbf{V}||_2^2 \end{aligned} \tag{8}$$

Based on the above two objectives, we formulate the learning process of topic embedding as a joint-objective optimization. We model the objective function as a linear combination of the above two objectives

$$J_U(\theta) = \gamma \cdot J_T(\theta) + (1-\gamma)\cdot J_C(\theta) \tag{9}$$

where $\gamma \in [0,1]$ is a preference parameter. The optimal solution of parameters is

$$\theta^* = \arg\min_\theta J_U(\theta) \tag{10}$$

Since the size of $\mathbf{E}$ in Equation 5 is $\frac{|T|(|T|-1)}{2}$, calculating the normalization part is quite time consuming. To address this challenge, we use the Noise Contrastive Estimation (NCE) [11] to estimate the parameters in our objective function. NCE provides a principle for unnormalized statistical models, which has been applied in estimating language models, word embedding, and anomaly detection [7, 20, 21]. NCE considers the normalization constant as an additional parameter of the model. We first consider the normalization constant as a parameter c. The probability in Equation 5 is thus re-written as:

$$P_\theta(e) = \exp(S_{\theta_0}(e) + c) \tag{11}$$

where $\theta = \{\theta_0, c\}$ represents the new parameters to be learned. In NCE, artificially generated noise data is added to the training data, and both parameters in probability density function and normalization constant can be estimated by discriminating the original data and noise data. The artificial noise distribution, denoted by $P_n(e)$, is the probability of an event e to be a noise sample. For each observed event e, we sample k noise samples $\{e'\}$ according to P_n. As for the chosen of P_n, it can be some factorized distribution on the event space, which can be specified uniformly or computed by counting frequency of topics in dataset. In this paper we use the strategy of counting frequency as it has been reported to be better [7]. We use $D = 1$ to indicate the event e in the observed data set $\mathbf{E}$ and $D = 0$ to indicate an event from the noise sample. The posterior probability is:

$$\begin{aligned} P(D=1|e,\theta) &= \frac{P_\theta(e)}{P_\theta(e) + kP_n(e)} \\ &= \sigma\Big(\log P_\theta(e) - \log kP_n(e)\Big) \end{aligned} \tag{12}$$

$$\begin{aligned} P(D=0|e,\theta) &= \frac{kP_n(e)}{P_\theta(e) + kP_n(e)} \\ &= 1-\sigma\Big(\log P_\theta(e) - \log kP_n(e)\Big) \end{aligned} \tag{13}$$

where $\sigma(x) = \frac{1}{1+\exp(-x)}$ is the sigmoid function. Now we fit the model by maximizing the expectation of log-posterior probability over the mixture of observed samples and noise samples. The expectation is formulated as follows:

$$\begin{aligned}&E_{P_\theta}[\log P(D=1|e,\theta)]+\\&kE_{P_n}[\log P(D=0|e,\theta)]\\=&E_{P_\theta}\left[\log\sigma\Big(\log P_\theta(e)-\log kP_n(e)\Big)\right]+\\&kE_{P_n}\left[\log\Big(1-\sigma\Big(\log P_\theta(e)-\log kP_n(e)\Big)\Big)\right]\end{aligned}\tag{14}$$

Then the loss function of an event and its noise samples is formulated as:

$$\begin{aligned}J_T(\theta)=&-\log\sigma\Big(\log P_\theta(e)-\log kP_n(e)\Big)-\\&\sum_{e'}\log\Big(1-\sigma\Big(\log P_\theta(e')-\log kP_n(e')\Big)\Big)\end{aligned}\tag{15}$$

The gradient function for $\mathbf{V}$ in $J_T(\theta)$ is

$$\begin{aligned}\frac{\partial J_T(\theta)}{\partial\mathbf{V}}=&\left[\sigma\Big(\log P_\theta(e)-\log kP_n(e)\Big)-1\right]\frac{\partial S_\theta(e)}{\partial\mathbf{V}}\\&+\sum_{e'}\left[\sigma\Big(\log P_\theta(e')-\log kP_n(e')\Big)\right]\frac{\partial S_\theta(e')}{\partial\mathbf{V}}\end{aligned}\tag{16}$$

Since the gradient function for c is similar to $\mathbf{V}$, for presentation simplification, we do not present it in this paper. The gradient function for another objective function $J_C(\theta)$ is formulated as follows:

$$\frac{\partial J_C(\theta)}{\partial\mathbf{V}}=\sum_{u\in U_{\text{seed}},v\in U,u\neq v}\left[\mathbf{d}_u(\mathbf{d}_u'^{\top}-\mathbf{d}_v'^{\top})+(\mathbf{d}_u'-\mathbf{d}_v')\mathbf{d}_u^{\top}\right]\mathbf{V}\tag{17}$$

Algorithm 1: Adam SGD Algorithm

```
Require: $\alpha$: Stepsize
Require: $\beta_1,\beta_2\in[0,1)$: Exponential decay rates for the
  moment estimates
Require: $f(\theta)$: Stochastic objective function with
  parameters $\theta$
Require: $\theta_0$: Initial parameter vector
$m_0\leftarrow 0,\ v_0\leftarrow 0,\ t\leftarrow 0$
while $\theta_t$ not converged do
    $t\leftarrow t+1$
    $g_t\leftarrow\nabla_\theta f_t(\theta_{t-1})$(Get gradients w.r.t stochastic objective
      at timestep $t$)
    $m_t\leftarrow\beta_1\cdot m_{t-1}+(1-\beta_1)\cdot g_t$
    $v_t\leftarrow\beta_2\cdot v_{t-1}+(1-\beta_2)\cdot g_t^2$
    $\widehat{m}_t\leftarrow m_t/(1-\beta_1^t)$
    $\widehat{v}_t\leftarrow v_t/(1-\beta_2^t)$
    $\theta_t\leftarrow\theta_{t-1}-\alpha\cdot\widehat{m}_t/(\sqrt{\widehat{v}_t}+\varepsilon)$
end
return $\theta_t$ (Resulting parameters)
```

Learning Algorithm

In our approach, we adopt the Stochastic Gradient Decent (SGD) method for learning the parameters. To speed up the learning procedure, we propose a weighted joint-objective optimization algorithm based on Adam [14].

The Adam algorithm has been shown to work well in practice and to favorably compare to other adaptive learning methods. To make this paper self-contained, we present the Adam steps in Algorithm 1. Our algorithm is summarized in Algorithm 2. In each iteration, for efficient computation, we randomly select an objective to update parameters $\mathbf{V}$ based on γ. And we sample a mini-batch of topic co-occurrences for objective J_T and sample a user from seed user set for objective J_C. All parameters in Adam, except for $\mathbf{V}$, are independent from each other in the optimization process of the two objectives.

Algorithm 2: Adam based Joint-objective SGD Algorithm

```
Input: Size of mini-batch $n$; Preference weight $\gamma$; Training
       samples of topic co-occurrences $\mathbf{E}$; Users set $U$ and
       seed user set $U_{\text{seed}}$
Output: Embedding matrix $V$
Initialize $V$ randomly
while $J_U(\theta)$ not convergenced do
    $x=random(0,1)$ //generate a real number $x\in[0,1)$
    if $x<\gamma$ then
        Sample $n$ topic co-occurrence events $e$ and $k*n$ noise
          events $e'$ randomly
        Select $J_T(\theta)$ as objective function, perform an
          iteration in Adam to update $\mathbf{V}$
    end
    else
        Sample a user $u_i$ randomly
        Select $J_C(\theta)$ as the objective function, perform an
          iteration in Adam to update $\mathbf{V}$
    end
end
return $\mathbf{V}$
```

EXPERIMENTS

Datasets

We use two real datasets, *MovieLens* and *Zhihu*, to experimentally evaluate the proposed method. The statistics of the two datasets are listed in Table 2. Details are given below.

MovieLens Dataset. The *MovieLens 20M* dataset released by Grouplens [13] contains user rating and free-text tagging activities on *MovieLens*, a popular movie recommendation platform. It contains data created by 138493 users between January 09, 1995 and March 31, 2015. Since most users do not keep active across 20 years, we select a part of the dataset which is created from April 2009 to March 2015, in which enough active users exist during this period. Users' ratings on movies are considered as user behavior records. Each pair of movie and tag is associated with a relevance score ranging from 0 to 1, using the Tag Genome approach [28]. Considering the semantic correlation between tags and a movie, only tags with relevance scores larger than 0.7 are selected as topics on this movie. To evaluate our method for solving the UIL problem, we partition the selected dataset into two parts according to time period: the first part $\mathbb{B}_1$ covers ratings from April 2009

Table 2. Statistics of Experiment Datasets

Datasets	# users	# topics	# events	# records
Zhihu	1,861	2,590	2,710,804	2,935,482
MovieLens	1,857	1,100	2,396,979	831,106

to April 2012, and second part $\mathbb{B}_2$ covers ratings from April 2012 to April 2015.

Zhihu Dataset. *Zhihu* is a Chinese Q&A website where questions are created, answered, edited and organized by the platform audience. We crawled user behavior records of around 8,000 users from October 2015 to September 2016, such as answering questions, voting up answers, etc. The tags marked on questions are selected as topics. The dataset is also partitioned into two parts: the first part $\mathbb{B}_1$ covers behaviors from October 2015 to March 2016, and second part $\mathbb{B}_2$ covers behavior data made from April 2016 to September 2016.

Evaluation Metrics and Distance Metrics

For the UIL problem, a widely adopted evaluation metric is to calculate the top-k similar candidates for a target user and verify whether the true identity is within the results. In our setting, for each user $u \in \mathbb{B}_2$, we calculate its distances with users in $\mathbb{B}_1$ and rank them in an ascending order. The index function $hit(u)$ is used to verify whether user u is correctly mapped to the same identity in $\mathbb{B}_1$ within the top-k users. $hit(u) = 1$ indicates that u has been correctly linked, $hit(u) = 0$ otherwise. Let U_{test} denote the set of test users, the accuracy for identity linkage is defined as follows:

$$acc = \frac{\sum_{u \in U_{test}} hit(u)}{|U_{test}|} \tag{18}$$

There are many candidate distance functions. Considering the semantics of user vectors, we adopt the Cosine distance and Euclidean distance in most experiments. Since some comparison methods adopt the probability distribution as user vectors, we also adopt the balanced KL divergence as the distance metrics in this method. For example, Naini et al. adopt the balanced KL divergence for user matching problem. It performs well in their statistics method [22, 27]. For justification purpose, we adopt their best results for comparison. However, this metric is not suitable for the intrinsic characteristic vectors that we use in our approach. So this metric is only used in the comparison methods.

Another evaluation strategy is matching all users simultaneously, which can be seen as the problem of minimum-weight perfect matching on a bipartite graph. We do not choose this strategy for two reasons. One is the complexity of the problem. The well-known Hungarian algorithm can solve the perfect matching problem with complexity $O(n^3)$ [17] where n represents the number of users. As n increases, the computation cost becomes high. Another reason is that the *exact matching* definition is not common in practice.

Comparison methods and settings

We compare the following state-of-art methods for UIL.

Statistics. In this method, the probability distribution over topics is seen as a user characteristic vector [22, 27], which is directly applied to the distance between users.

NMF. We also consider the topic model NMF [26], because it is similar to our method from the point of view of discovering latent characteristics. We define hyper-topic (or latent factor) as a higher level generalization among topics. We use the user-topic probability distribution matrix as the input of NMF. NMF outputs the relevance between user and hyper-topics, which is used as the user vectors for identity linkage.

E-T (Embedding with Topic compatibility). This method is a specific form of our proposed method which learns the embedding using only the objective of topic compatibility and ability to distinguish users, namely γ is equal to 1.

E-C (Embedding with Characteristic consistency). This method is the second form of our proposed method which learns the embedding using only the objective of user intrinsic characteristic consistency. It means that γ is set to 0.

E-TC (Embedding with Topic compatibility and Characteristic consistency). This is the general form of our proposed method, which learns the embedding using a joint-objective optimization.

The dimension of the latent space in our methods and the numbers of hyper-topics in NMF are set to 50. All parameters to be learned are initialized randomly. For each observed topic co-occurrence, we draw one negative sample. To speed up the computation of the stochastic gradient descent, we set a mini-batch of 1024 for sampling topic co-occurrence events. For the experiments on the *Zhihu* dataset, we set a behavior threshold of 100, which means we choose users who have more than 100 behavior records in both periods. We explore how this threshold influences identity linkage in the experiments. We also filter out topics that occur less than 200 times. Other parameters are set as follows: $\varepsilon = 10^{-1}, \gamma = 0.5, \lambda = 0$. In *MovieLens* dataset, we set behavior threshold to be 50 and $\varepsilon = 10^{-2}, \gamma = 0.1, \lambda = 10^{-5}$. We adopt a 5-fold cross-validation in our experiments.

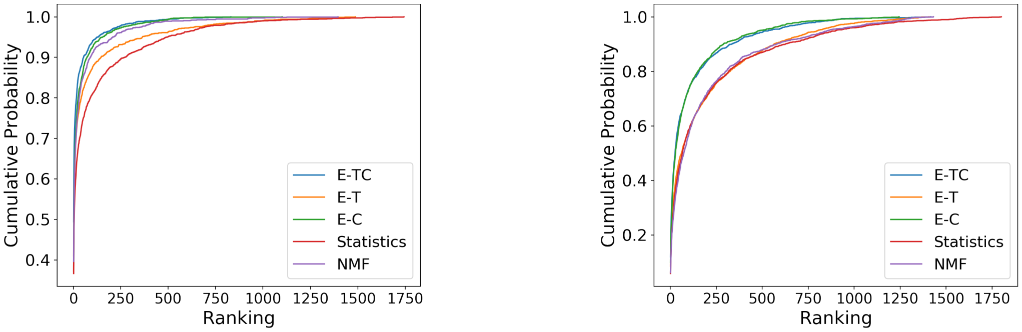
Results for Identity Linkage

We first verify the accuracy of our methods and other comparison methods. The results, reported in Table 3, show that method **E-TC** largely outperforms other methods. On top-1 identity linkage, the accuracy of **E-TC** reaches nearly 50% in *Zhihu* dataset and nearly 13% in *MovieLens* dataset, respectively, which is 12% and 6% higher than the method NMF. On top-5 and top-10 linkage, the accuracy of method **E-TC** outperforms on both datasets. Furthermore, in the *MovieLens* dataset, an increasing k enhances the accuracy of method **E-TC** compared with the other methods. This good performance shows that although the learned characteristics of users can not guarantee an exactly identity matching, they are able to provide good approximations for recognizing a user.

When we use only a single objective to learn the embedding, namely methods **E-T** or **E-C**, the accuracy is still higher than the accuracy of baseline methods on both datasets. It is interesting to notice that in the *Zhihu* dataset under the Cosine

Table 3. Accuracy of Identity Linkage Compared with Different Methods

	top-1			top-5			top-10		
Distance Metric	cos	Euc	KL	cos	Euc	KL	cos	Euc	KL
Zhihu Dataset									
Statistics	0.366	0.302	0.354	0.528	0.435	0.493	0.586	0.489	0.564
NMF	0.379	0.344		0.576	0.520		0.651	0.584	
E-T	0.419	0.441		0.603	0.623		0.672	0.695	
E-C	0.372	0.372		0.607	0.602		0.693	0.687	
E-TC	**0.497**	**0.469**		**0.695**	**0.664**		**0.762**	**0.730**	
MovieLens Dataset									
Statistics	0.059	0.055	0.043	0.162	0.142	0.113	0.229	0.209	0.153
NMF	0.065	0.058		0.156	0.143		0.248	0.219	
E-T	0.090	0.088		0.210	0.208		0.280	0.277	
E-C	0.104	0.100		0.246	0.237		0.336	0.322	
E-TC	**0.129**	**0.126**		**0.279**	**0.270**		**0.366**	**0.355**	



(a) CDF of ranking in Zhihu

(b) CDF of ranking in MovieLens

Figure 3. The CDF of ranking of users with themselves in different methods. The X-axis corresponds to the ranking of true identity linkage. The Y-axis represents the cumulative distribution of users, which means the proportion of users whose ranking is less than the value in x-axis. Both figures indicate that introducing seed users (E-TC, E-C) improves the learning of the embedded user behavior semantics.

distance, method **E-T** outperforms **E-C** in the case of top-1 linkage. But as k increases, method **E-C** gradually outperforms **E-T**. Such a result indicates that the objective of topic compatibility is quite helpful when performing accurate identity linkage, and the objective of characteristic consistency makes a user's trace more identical from the global perspective.

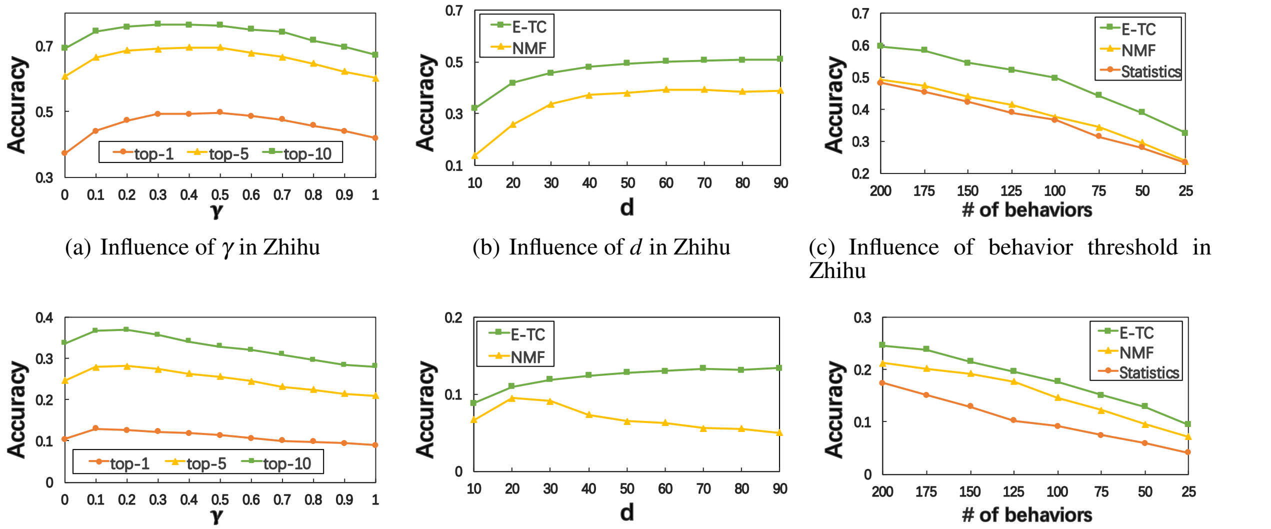
When comparing different distance metrics, there is no obvious difference between Cosine and Euclidean distances. In most cases, the Cosine distance performs slightly better than the Euclidean distance except in the *Zhihu* dataset by method **E-T**. So we adopt the Cosine distance metric in the rest of the experiments.

Although our methods can not exactly link every user identity, they do reduce the difficulty in recognizing a specific user from a large user set. To have a clear explanation of such a result, for each anonymized user $u \in \mathbb{B}_2$, we calculate the distance between u and every user $v \in \mathbb{B}_1$, and count the ranking of his or her true identity. The smaller the ranking, the better the linkage performance. Figure 3 shows the Cumulative Distribution Function(CDF) curve on the rankings for the whole test dataset. We can see that, in both datasets, method **E-TC** performs the best, shown as the fast rising curve, followed tightly by method **E-C**. Such results show that the introducing seed users provides more background knowledge; thus the embedding method can learn more comprehensive semantics from user behaviors. Since method **E-T** does not introduce such background knowledge, it performs worse than the other two.

Influence of Parameters and Settings

We first analyze the preference parameter γ, which is the trade-off term for two learning objectives. Figure 4(a) shows the performance for different values of γ on the *Zhihu* dataset, where $\gamma = 0$ indicates using only the characteristic consistency objective, and $\gamma = 1$ indicates using only the topic compatibility objective. We can see that when γ is around 0.3-0.4, our method performs best. Figure 4(d) shows the performance for different values of γ in the *MovieLens* dataset. Our method performs best when $\gamma = 0.1$. Such results show that, in the *MovieLens* dataset, tags are not so semantically relevant as in the *Zhihu* dataset. Consequently, the results show the small importance of topic compatibility in optimization.

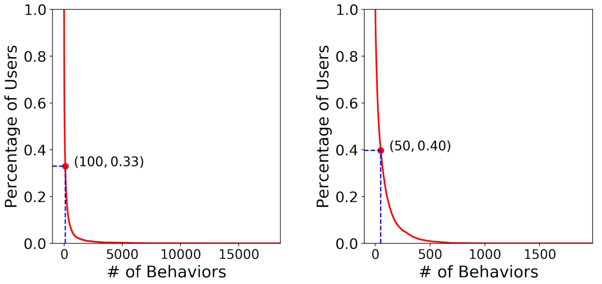
Then we analyze the impact of the embedding dimension parameter d. As a comparison, we choose the same dimension



(a) Influence of γ in Zhihu (b) Influence of d in Zhihu (c) Influence of behavior threshold in Zhihu

(d) Influence of γ in MovieLens (e) Influence of d in MovieLens (f) Influence of behavior threshold in MovieLens

Figure 4. Influence of parameters and settings and comparison with other methods on two datasets. Figure (a)(d) show the performance with respect to γ. Figure (b)(e) show the performance with respect to d. Figure (c)(f) show the performance with respect to behavior threshold.The optimal value that leads to the best performance is dependent on the dataset for all parameters. Our method E-TC outperforms the other methods consistently.



(a) # of behaviors in Zhihu (b) # of behaviors in MovieLens

Figure 5. Statistics on user behavior quantity. X-axis represents the number of behaviors, and Y-axis represents the proportion of users whose number of behaviors is higher than a given threshold in both periods. We can see that users in both datasets follow a long-tail distribution. The selected behavior threshold in this paper is marked by the red point.

Table 4. Selected topics in three themes

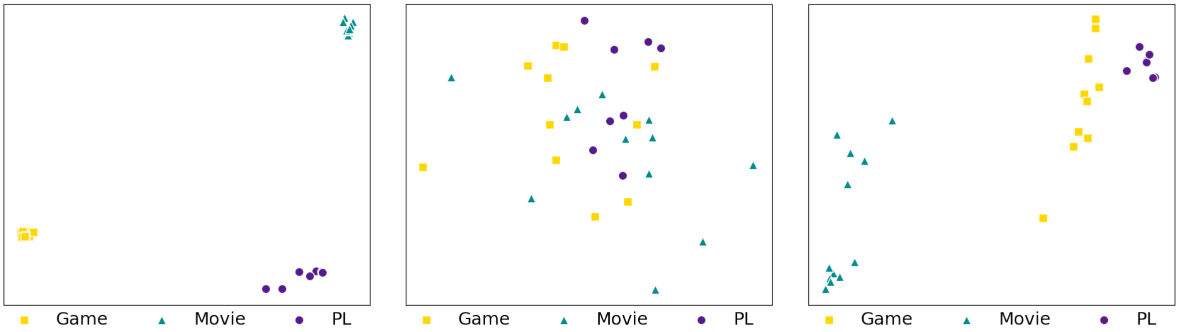
Theme	Topics
Game	Overwatch, LOL, Dota2, Hearthstone Dota, MOBA, Clash of Clans, Steam Minecraft, iOS Game, Game Design
Movie	Hongkong film, micro film, Douban film Chinese film, American film, horror film Japanese film, film, Hollywood film Korea film, science fiction film
Programming Language	c/c++, JavaScript, JAVA, Python PHP, C#, Node.js, C

as the number of hyper-topics in method NMF. Figures 4(b) and 4(e) show the performance with respect to d in the two datasets. We can see that in the *Zhihu* dataset, the accuracy increases fast for both methods when d is less than 50 and becomes stable after d reaches 50. In the *MovieLens* dataset for method **E-TC**, the accuracy increases fast when d is smaller than 40. But for method NMF, the peak accuracy is at about $d = 20$. Since a larger d means more computation, in practice, we should take into account both accuracy and computation cost. In most of the experiments reported this paper, we choose d as 50 and 40 for the two datasets.

We also analyze how the number of user behaviors influences identity linkage. Figures 5(a) and 5(b) show the statistics about the number of behaviors in the two datasets, where the X-axis represents the number of behaviors, and the Y-axis represents the proportion of users whose number of behaviors is higher than a given threshold in both periods. We can see that the number of user behaviors in both datasets follow a long-tail distribution. A large amount of users have only a small amount of behaviors. We conduct our experiments on different settings for the behavior threshold; the results are shown in Figures 4(c) and 4(f). As the behavior threshold decreases, the accuracy decreases since a large amount of users with a small amount of behaviors are taken into account. This makes it difficult to learn user intrinsic characteristics from a small quantity of behaviors. However, our method **E-TC** outperforms the other methods consistently.

Understanding Semantics of Topic Embedding

To provide more understandable results for other applications, we evaluate the semantics of embedding from two perspectives: topic relevance and user intrinsic characteristics. To give an intuitive view on the meanings of topic embedding, we first use the t-Distributed Stochastic Neighbor Embedding (t-SNE) technique [19] to find 2d coordinates of the original



(a) Embedding by E-T (b) Embedding by E-C (c) Embedding by E-TC

Figure 6. The 2d coordinates representations of the embedding of selected topics for three methods. Topics selected from the theme of *Game*, *Movie* and *Programming Language* are labeled by yellow square, green triangle, purple solid circle, respectively. We can see that the results of dimensionality reduction of three methods are quite different.

Table 5. Top 10 topics concerned by three users in two time periods

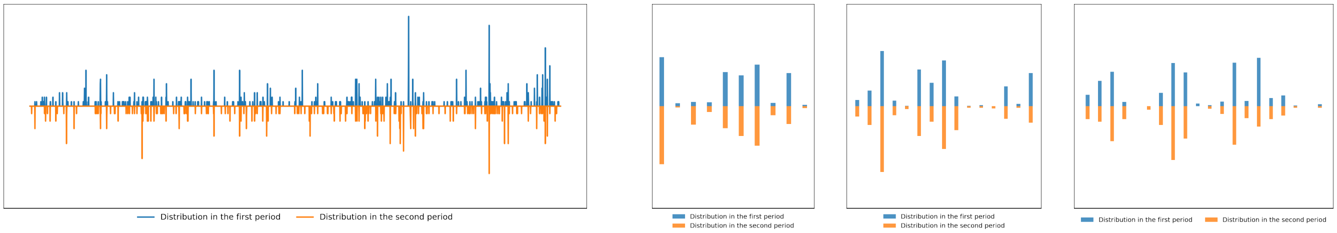
	In the first time period	In the second time period	Ranking promotion
User 1	Living, Internet, History **Artificial Intelligence**, Society Experience, **Deep Learning** **Google**, **Computer**, **Baidu**	Internet, **Math**, Health **Alibaba**, **Technology** **Didi Travel(online hailed car)** Law, History,**Google**, **Travel O2O**	1200
User 2	Living, Literature, Experience Psychology, Film, Novel Interpersonal communication **Programmer**, society,survey question	**Python**, **Program**, Living Medical Science, Health, **Programmer** **Crawler(Computer Networks)** Life, Life history, Internet	912
User 3	Living, **Psychology**, Experience Homosexual, Variety Show, **Art** **Photograph**, **Mentality**, Film survey question	**Design**, Japan, Living **Drawing**, **Art**, **Graphic Design** **Psychology**, **Photoshop** **Photograph**, Film	652

topic embedding in the *Zhihu* dataset. t-SNE is a technique for dimensionality reduction particularly well suited for the visualization of high-dimensional datasets. We select topics from 3 themes, which are *Game*, *Movie* and *Programming Language(PL)*, respectively. The details of selected topics are reported in Table 4. We color topics according to their themes. We expect points in the same color to be clustered together, and each point is be distinguished from others.

Figure 6(a) shows the embedding learned by method **E-T**. We can see that the majority of topics in the same theme are clustered quite closely. There are clear boundaries between clusters of different themes. But topics in the same theme cannot be distinguished from each other. The reason is that in method **E-T** we learn semantics of topics based on their compatibilities, which is modeled based on the co-occurrences of topics. Our objective is to let the embedding of co-occurred topics be as close as possible. However, each topic has its own semantic which cannot be completely the same with similar topics. Using only the information of topic co-occurrence seems not enough to learn topic semantics comprehensively. Figure 6(b) shows the topic embedding learned by **E-C**. We can see that topics are roughly clustered together without clear boundaries between them. The reason is that method **E-C** considers user characteristic consistency. The topics of interest to the same user are learned to be close in their vectors. Since this method uses the seed users rather than the platform population, it may lead to showing irrelevant topics to be close due to some users' occasional activities.

The semantics learned by **E-TC** overcomes the shortcomings of methods **E-T** and **E-C**. Figure 6(c) shows the topic embedding learned by **E-TC**. We can see that topics in the same theme are clustered together and the boundaries between clusters are obvious. It is worth noting that clusters for *Game* and *Programming Language* are closer than the cluster *movie* in method **E-TC**. Such a result reflects the fact that the background knowledge about seed users reveals that programmers are more likely to enjoy games, which is not shown by results obtained by method **E-T**.

Then we try to understand how the semantics solve UIL compared to other method. We analyze the representative users who are well recognized with great ranking promotion by our method than the Statistics method. We choose three such users from the *Zhihu* dataset and list the top-10 topics of interest for each one in Table 5. Although some common topics such as *Living*, *Experience*, *Life* and *Internet* are the same, most



(a) original topics (b) 10 clusters (c) 15 clusters (d) 20 clusters

Figure 7. A user's behaviors in two time periods are described in four forms of probability distributions over: (a) original topics; (b) 10 clusters; (c) 15 clusters; and (d) 20 clusters. The blue line denotes the first period, the orange line denotes the second period, and the clusters are classified based on learned topic embedding representations.

topics of interest for the same user in the two periods vary a lot. That is why the Statistics method cannot recognize them correctly. But we can catch some intrinsic characteristics from the semantic related topics (highlighted by colors in the table). For example, the first user is probably employed in an Internet company according to the topics in red, and the blue topics show his/her interests in the domain of artificial intelligence. The second user is probably a network programmer according to the topics colored in blue. For the third user, the blue topics indicate that he or she is a graphic designer, and red topics indicate the interests in psychology.

To further understand how the semantics of embedding help recognize a user, we compare a user's behavior pattern in different modes. First, we cluster all topics into k classes based on their embedding representations and then represent each user as the probability distributions over these classes. An example is given in Table 7. For the above first user, his or her behaviors are modeled as the probability distributions over topics and clusters, for $k = 10$, $k = 15$ and $k = 20$, respectively. The results show that the variants on user's topics of interest between two periods have been highly reduced under the learned embedding representation, and they are similarly consistent on different k settings. As we expected, it indicates that the embedding method learned the consistent intrinsic characteristics implied in user behaviors.

CONCLUSION

In this paper, we solve the problem of user identity linkage on social media by discovering user intrinsic characteristics. We propose an embedding method to understand the semantics of topics related to user behaviors. The embedding representations of topics are learned by a joint-objective optimization, which tries to maximize the topic compatibility, discriminating ability, and characteristic consistency of the seed user. Experimental results on two real social media datasets show that our method outperforms other related methods. We also analyze the semantics of embedding representations from both topic view and user behavior perspective. The effectiveness of our method highlights the privacy and identity issues in online communities. It reminds users to harness their behaviors to avoid being de-anonymized. From another perspective, for online service providers, they should provide more protection on users' private information.

As for future work, we plan to investigate the identity linkage problem on different social media platforms. Since users may have different interests and behave differently on different platforms, understanding topics from different sources and embedding them into the same latent space will be more complex and challenging. Another interesting direction is the selection of seed users. As we discussed in the paper, seed users allows us to introduce background knowledge that helps better understand the intrinsic characteristics of users. We would like to investigate how to select seed users to improve the performance under a given computation and seeding budget.

ACKNOWLEDGMENTS

We would like to thank Gong Hao from Fudan University, Shaoqing Wang from Shandong University for many helpful discussions. This work is supported by the National Natural Science Foundation of China (91646119), Shandong Provincial Natural Science Foundation under Grant No. ZR2014FM014, the Key Research and Development Program of Shandong Province (2014GGX101046, 2015GGX106002, 2017GGX10114), the Special Program on Independent Innovation & Achievements Transformation of Shandong Province (2014ZZCX03301) and SAICT Expert Program.

REFERENCES

1. Einat Amitay, Sivan Yogev, and Elad Yom-Tov. 2007. Serial sharers: Detecting split identities of Web authors. In *Proceedings of the SIGIR 2007 International Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection PAN 2007, Amsterdam, Netherlands, July 27, 2007.*

2. Saeideh Bakhshi, Partha Kanuparthy, and David A. Shamma. 2015. Understanding Online Reviews: Funny, Cool or Useful?. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work &#38; Social Computing*. ACM, New York, NY, USA, 1270–1276.

3. Sergey Bartunov, Anton Korshunov, Seung-Taek Park, Wonho Ryu, and Hyungdong Lee. 2012. Joint link-attribute user identity resolution in online social networks. In *Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining, Workshop on Social Network Mining and Analysis. ACM*.

4. Mikhail Belkin and Partha Niyogi. 2002. Laplacian eigenmaps and spectral techniques for embedding and

clustering. In *Advances in neural information processing systems*. 585–591.

5. David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.
6. Ji-Won Byun, Ashish Kamra, Elisa Bertino, and Ninghui Li. 2007. Efficient k-anonymization using clustering techniques. In *International Conference on Database Systems for Advanced Applications*. Springer, 188–200.
7. Ting Chen, Lu-An Tang, Yizhou Sun, Zhengzhang Chen, and Kai Zhang. 2016. Entity embedding-based anomaly detection for heterogeneous categorical events. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*.
8. Yoni De Mulder, George Danezis, Lejla Batina, and Bart Preneel. 2008. Identification via location-profiling in GSM networks. In *Proceedings of the 7th ACM workshop on Privacy in the electronic society*. ACM, 23–32.
9. Sebastien Gambs, Marc-Olivier Killijian, and Miguel Nunez del Prado Cortez. 2014. De-anonymization attack on geolocated data. *J. Comput. System Sci.* 80, 8 (2014), 1597–1614.
10. Hasini Gunasinghe and Elisa Bertino. 2016. RahasNym: Pseudonymous Identity Management System for Protecting against Linkability. In *Collaboration and Internet Computing (CIC), 2016 IEEE 2nd International Conference on*. IEEE, 74–85.
11. Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models.. In *AISTATS*, Vol. 1. 6.
12. Jiawei Han, Jian Pei, and Micheline Kamber. 2011. *Data mining: concepts and techniques*. Elsevier.
13. F Maxwell Harper and Joseph A Konstan. 2016. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 5, 4 (2016), 19.
14. Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*.
15. Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009).
16. Nitish Korula and Silvio Lattanzi. 2014. An efficient reconciliation algorithm for social networks. *Proceedings of the VLDB Endowment* 7, 5 (2014), 377–388.
17. Harold W Kuhn. 2010. The hungarian method for the assignment problem. *50 Years of Integer Programming 1958-2008* (2010), 29–47.
18. Siyuan Liu, Shuhui Wang, and Feida Zhu. 2015. Structured Learning from Heterogeneous Behavior for Social Identity Linkage. *IEEE Transactions on Knowledge and Data Engineering* 27, 7 (2015), 2005–2019.
19. Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, Nov (2008), 2579–2605.
20. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations*.
21. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.
22. Farid M Naini, Jayakrishnan Unnikrishnan, Patrick Thiran, and Martin Vetterli. 2016. Where you are is who you are: User identification by matching statistics. *IEEE Transactions on Information Forensics and Security* 11, 2 (2016), 358–372.
23. Arvind Narayanan and Vitaly Shmatikov. 2008. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*. IEEE, 111–125.
24. Arvind Narayanan and Vitaly Shmatikov. 2009. De-anonymizing social networks. In *2009 30th IEEE symposium on security and privacy*. IEEE, 173–187.
25. Olga Peled, Michael Fire, Lior Rokach, and Yuval Elovici. 2013. Entity matching in online social networks. In *Social Computing (SocialCom), 2013 International Conference on*. IEEE, 339–344.
26. Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. 2012. Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, 952–961.
27. Jayakrishnan Unnikrishnan and Farid Movahedi Naini. 2013. De-anonymizing private data by matching statistics. In *Communication, Control, and Computing (Allerton), 2013 51st Annual Allerton Conference on*. IEEE, 1616–1623.
28. Jesse Vig, Shilad Sen, and John Riedl. 2012. The tag genome: Encoding community knowledge to support novel interaction. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 2, 3 (2012), 13.
29. Hui Zang and Jean Bolot. 2011. Anonymization of location data does not work: A large-scale measurement study. In *Proceedings of the 17th annual international conference on Mobile computing and networking*. ACM, 145–156.