User Identity Linkage across Social Media via Attentive Time-aware User Modeling

Xiaolin Chen, Xuemeng Song, Siwei Cui, Tian Gan, Zhiyong Cheng, and Liqiang Nie, *Member, IEEE*

***Abstract*—In this paper, we work towards linking users' identities on different social media platforms by exploring the user-generated contents (UGCs). This task is non-trivial due to the following challenges. 1) As UGCs involve multiple modalities (e.g., text and image), how to accurately characterize the user account based on their heterogeneous multi-modal UGCs poses the main challenge. 2) As people tend to post similar UGCs on different social media platforms during the same period, how to effectively model the temporal post correlation is a crucial challenge. And 3) no public benchmark dataset is available to support our user identity linkage based on heterogeneous UGCs with timestamps. Towards this end, we present an attentive time-aware user identity linkage scheme, which seamlessly integrates the temporal post correlation modeling and attentive user similarity modeling. To facilitate the evaluation, we create a comprehensive large-scale user identity linkage dataset from two popular social media platforms: Instagram and Twitter. Extensive experiments have been conducted on our dataset and the results verify the effectiveness of the proposed scheme. As a residual product, we have released the dataset, codes, and parameters to facilitate other researchers.**

***Index Terms*—Attention Mechanism, User Identity Linkage, Temporal Correlation.**

I. Introduction

RECENT years have witnessed a growing trend among people to immerse themselves in multiple social media platforms and hence fully enjoy the various services. According to the report by Pew Research Center in 2018[1], more than three-quarters of Internet users are using two or more social media platforms. In particular, 73% of Twitter users are involved in Instagram concurrently, and 60% of Instagram users also visit Snapchat. Such a phenomenon has inspired researchers to investigate tasks across social media [1–7], ranging from the user migration [1] to the personalized recommendation [2, 3, 5, 6]. As the key prerequisite of existing studies, the research of user identity linkage which is to link individual's accounts on different social media platforms have received increasing attention.

As a matter of fact, several efforts have been dedicated to solving the problem of user identity linkage [8–11]. They mainly focus on investigating the user profiles (e.g., username, birthday and gender) [8, 12] and social connection structures [9, 11]. Despite the great success achieved by these efforts, most of them overlook the UGCs, whereby the information is richer and the cues conveying users' characteristics are more representative. Although some pioneer studies [13] have attempted to bridge this gap, they mainly utilize the shallow learning methods and focus on the homogeneous UGCs (e.g., the text), resulting in suboptimal in practical applications. Meanwhile, recent years have witnessed the compelling advances of deep neural networks in representation learning [14], which propels us to explore the potential of incorporating the deep learning technique in the context of user identity linkage, especially using the heterogeneous UGCs (e.g., the text and image).

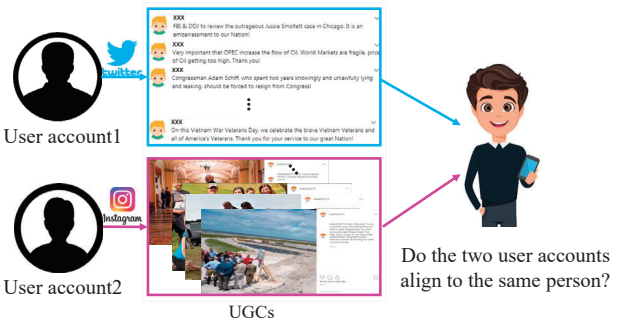

Fig. 1. Illustration of the user identity linkage task.

In this work, we aim to study the practical problem of user identity linkage across social media by particularly answering the question *"whether the given user accounts on different social media refer to the same identity based on their heterogeneous UGCs*", as illustrated in Figure 1. The task we propose here primarily involves the user similarity modeling across different social media platforms. However, the user similarity modeling based on heterogeneous UGCs is non-trivial due to the following reasons. 1) UGCs may involve multiple modalities, including texts, images, and videos. Intuitively, the image and corresponding textual description in a UGC may have different confidence pertaining to the user characterization. Figure 2 shows two UGC examples. It is apparent that the textual description delivers more signals in characterizing the user in the first UGC, while the image is more informative in the second one. Therefore, how to adaptively characterize the confidence of different modalities towards user identity linkage poses the main challenge for us. 2) In fact, one user tends to post similar, even the same, UGCs across different social media in a short period. For example,

X. Chen, X. Song, T. Gan and L. Nie are with the School of Computer Science and Technology, Shandong University, Qingdao 266237, China (e-mail: cxlicd@gmail.com, sxmustc@gmail.com, gantian@sdu.edu.cn, nieliqiang@gmail.com).

S. Cui is with the Department of Computer Science and Engineering, Texas A&M University, College Station, TX 77843 USA (e-mail: siweicui@tamu.edu).

Z. Cheng is with Qilu University of Technology (Shandong Academy of Sciences), Jinan 250014, China (e-mail: jason.zy.cheng@gmail.com).

[1]https://tinyurl.com/y3htxhlq.

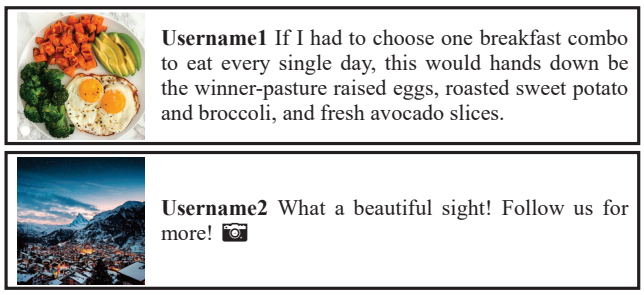

Fig. 2. UGCs with different modality confidence in user characterization.

a man may post the text "Looking forward to my trip to Beijing!!" on Twitter before his journey, while sharing pictures on Instagram during the trip. Consequently, how to incorporate the temporal post correlation between users' distributed UGCs into the user similarity modeling and thus conduct time-aware user identity linkage is a tough challenge. And 3) there is no publicly available large-scale dataset to well support our user identity linkage task based on users' heterogeneous UGCs with timestamps. Therefore, the last challenge lies in the lack of benchmark dataset.

To address the aforementioned challenges, we propose an attentive time-aware user identity linkage scheme, UserNet for short, as illustrated in Figure 3, which seamlessly integrates the *time-aware post correlation modeling* and *attentive user similarity modeling*. In particular, we define the time-aware post correlation modeling as incorporating the time factor to model the post correlation and hence identify the corresponding user accounts on different social networks. Essentially, we introduce a time decay factor to capture the pair-wise temporal correlation among posts, which intuitively reflects the influence of their post similarity on the final user similarity modeling. In addition, to accurately measure the user similarity, we propose the attentive user similarity modeling to adaptively fuse the user similarity compiled from different modalities. To comprehensively model the confidence of each modality regarding the user similarity modeling, the proposed model focuses on learning the global pair-wise post similarity distribution rather than the local post one. Moreover, to guarantee the evaluation quality and facilitate the experiment conduction, we construct a large-scale user identity linkage dataset, dubbed as TWIN, consisting of 5,765 users with 1,729,500 UGCs and corresponding timestamps on Twitter and Instagram. Extensive experiments on this real-world dataset have fully validated our work.

Our main contributions can be summarized in threefold:

- To the best of our knowledge, we are the first to incorporate the deep learning techniques in the context of the user identity linkage problem only based on the heterogeneous UGCs on different social media platforms. Previous work has either focused on the user profile and the social network structure or their combination with the user content. In particular, we present an attentive time-aware user identity linkage scheme, UserNet, which jointly unifies the *time-aware post correlation modeling* and *attentive user similarity modeling*.
- We propose the time-aware post correlation modeling to mine the temporal post correlation among one's distributed posts on different social media platforms. In addition, we further propose the attentive user similarity modeling that can adaptively fuse the user similarity compiled from different modalities.
- We create a comprehensive user identity linkage dataset, TWIN, consisting of 5,765 users with 1,729,500 UGCs and corresponding timestamps on Twitter and Instagram. Experimental results are encouraging with detailed analysis and insights. We have released our dataset, codes, and parameters to facilitate other researchers[2]. In particular, due to the concern of the privacy problem [15], we have obscured the sensitive user characteristics in our dataset.

II. Related Work

Both user identity linkage and representation learning are related to this work.

A. User Identity Linkage

As an emerging problem, user identity linkage has been arresting much attention recently [8–12, 16–18]. Initially, Fellegi et al. [19] first proposed the problem of record linkage which is the origin of the user identity linkage problem. Thereafter, Zafarani et al. [20] provided evidence on the existence of mappings among identities across multiple social media and first formally presented the user identity linkage problem. Besides, following the review [21], existing efforts can be broadly divided into three directions: user profile-based, network structure-based, and content-based approaches.

User profile-based approaches [8, 12] focus on investigating user profile attributes, such as the usernames, birthdays, and genders. For example, Perito et al. [8] investigated the feasibility of usernames to perform the user identity linkage with binary classifiers. In addition, Mu et al. [12] proposed a unified framework based on the latent user space modeling utilizing the basic user profiles. Despite their success, the user profile can be ambiguous and unreliable towards the user identity linkage, as it can be deliberately counterfeited by users.

Network structure-based approaches [9, 11] aim to link the user identities with their network structures. For example, Tong et al. [11] proposed a supervised network embedding model with the observed anchor links as the supervision to capture the major structural regularities. Zhou et al. [9] introduced the friend relationship-based user identification algorithm, which calculates a matching degree for all candidate User Matched Pairs (UMPs). UMPs with top ranks are considered to be identical. However, obtaining the complete social connection network of users can be intractable due to the large quantity and privacy issues, which hinders researchers from utilizing the graph structure patterns to link the user identities.

Content-based approaches [10, 13, 18, 22–27] attempt to link user identities by investigating the UGCs. For example, Rong et al. [10] developed an authorship identification framework, where the writing style features are exploited. Likewise,

[2]https://socialmediafpd.wixsite.com/usernet.

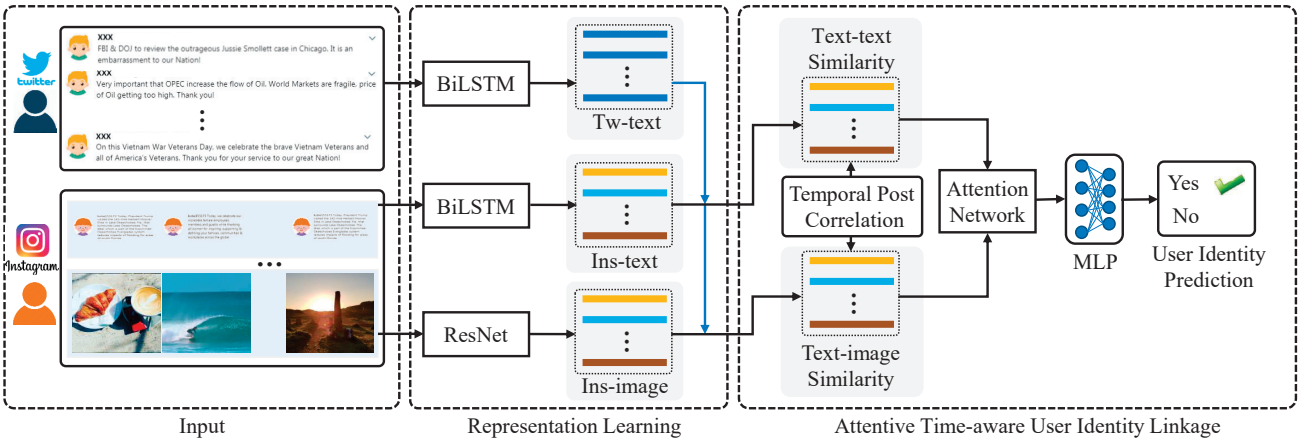

Fig. 3. Illustration of the proposed scheme for user identity linkage by exploring UGCs. In this scheme, we first learn the accurate representations for heterogeneous user posts on different social media platforms. We then conduct the attentive time-aware user similarity modeling to identify the same user. "Tw" and "Ins" refer to the Twitter and Instagram, respectively.

Olszewska et al. [27] presented a hausdorff-distance enhanced matching approach for image-to-image querying based on scale invariant feature transform descriptors. In addition, Goga et al. [18] proposed an n-gram language model to identify accounts on different social media platforms. Noticed the inherent correlation of the content, Sapumal et al. [13] introduced a method which enables the incorporation of user shared textual content as an approximate matching method.

In addition, several efforts have been made to explore multiple types of user information to boost the performance of user identity linkage [16, 28–33]. For example, Liu et al. [16] designed a heterogeneous behavior model to measure user behavior similarity from all aspects of a user's social data (i.e., profile, network and content). Besides, Ren [31] introduced a network alignment model, which predefines a set of meta diagrams to extract features of network structures and the content information.

Beyond the existing studies that mainly work on the homogeneous UGCs and overlooked the time factor, we aim to tackle the user identity linkage with the heterogeneous UGCs (e.g., the text and image) on different social media, where the temporal post correlation is also taken into account.

B. Representation Learning

As a hot research topic, representation learning has been striving for learning more precise representations of data, as compared to hand-crafted features, and hence achieves excellent performance in various tasks [34–38]. In particular, recently, the compelling success of deep neural networks has facilitated plenty of models, including Bi-directional Long Short-Term Memory (BiLSTM) [39], convolutional neural networks (CNN) [40] and deep Boltzmann machine [41]. For example, Wang et al. [42] employed deep auto-encoders to obtain the non-linear network structure and thus captured the accurate network embedding. Due to the increasingly complex data and tasks, multimodal learning has acquired some researchers' attention. For example, Ngiam et al. [43] introduced a structure to learn the shared representation of visual and speech data and thus resolved the problem of speech recognition. Karpathy et al. [44] studied multiple approaches to large-scale video classification using CNNs. Although representation learning has achieved compelling success in plenty of tasks, including multimedia retrieval [45], multilingual classification [46] and phonetic recognition [47], there is relatively sparse literature on user identity linkage.

III. PRELIMINARIES

We first introduce the necessary notations, and then formulate the research problem.

A. Notation

Formally, we first declare some notations. In particular, we use bold capital letters (e.g., $\mathbf{X}$) and bold lowercase letters (e.g., $\mathbf{x}$) to represent matrices and vectors, respectively. We employ non-bold letters (e.g., x) to denote scalars and Greek letters (e.g., γ) to parameters. If not clarified, all vectors are in column forms. The main notations used in this paper are summarized in Table I.

B. Problem Formulation

In this work, we aim to tackle the problem of user identity linkage across different social media platforms. Without loss of generality, we focus on linking user identities between two platforms O_1 and O_2, and it can be easily extended to the cases of multiple social media platforms. Moreover, we choose the popular social media Twitter and Instagram as O_1 and O_2 to illustrate our scheme. Owing to the different service emphases, people tend to broadcast news or join topic discussions via textual tweets on Twitter and share their daily life or thoughts by posting images with brief textual descriptions on Instagram [48]. To make the model more representative, we particularly assume that users prefer to generate mono-modal posts (e.g., text) on O_1 while multi-modal posts on (e.g., image-text pairs) on O_2.

TABLE I
SUMMARY OF THE MAIN NOTATIONS.

Notation	Explanation
$\mathcal{U}$	The set of training user account pairs.
$\mathbf{y}$	The label vector of training user account pairs.
$\mathcal{O}_1$	The first social media platform.
$\mathcal{O}_2$	The second social media platform.
u_1^i	The i-th user on $\mathcal{O}_1$.
u_2^i	The i-th user on $\mathcal{O}_2$.
$\mathcal{T}_i$	The related information of u_1^i.
$\mathcal{S}_i$	The related information of u_2^i.
t_k^i	The k-th textual post of u_1^i.
s_g^i	The g-th multi-modal post of u_2^i
$r_{k,g}$	The time decay factor between posts t_k^i and s_g^i.
$\hat{\mathbf{m}}_g$	The global similarity distribution of s_g^i towards t_k^i.
$\mathbf{d}$	The similarity representation between u_1^i and u_2^i.

Suppose we have a set of training user account pairs $\mathcal{U} = \{(u_1^1, u_2^1), (u_1^2, u_2^2), \cdots, (u_1^N, u_2^N)\}$, where each pair consists of two user accounts on social media O_1 and O_2, respectively. Meanwhile, we assume that for each user account u_1^i on O_1, we have his/her K historical mono-modal posts (i.e., text) with timestamps $\mathcal{T}_i = \{(t_k^i, p_k^i)\}_{k=1}^K$. Here t_k^i refers to the k-th textual post, while p_k^i denotes the corresponding timestamp. Analogously, for each user account u_2^i on O_2, we gather his/her G multi-modal posts (i.e., image-text pairs) $\mathcal{S}_i = \{(s_g^i, q_g^i)\}_{g=1}^G = \{(v_g^i, c_g^i, q_g^i)\}_{g=1}^G$, where v_g^i and c_g^i represent the visual and textual modality of the g-th post, respectively. q_g^i corresponds to the timestamp of the g-th post. The training user account pairs are labeled by $\mathbf{y} = (y_1, y_2, \cdots, y_N)^{\mathrm{T}} \in \mathbb{R}^N$, where y_i stands for the ground truth of the i-th pair (u_1^i, u_2^i). In particular, $y_i = 1$, if the accounts u_1^i and u_2^i refer to the same user identity in the physical world, and $y_i = 0$ otherwise. In a sense, we aim to learn the projection $\mathcal{G} : u_1^i \times u_2^i \to \{0, 1\}$ using the labeled training user account pairs, and on the top of that, we can identify whether u_1^i and u_2^i refer to the same identity.

IV. METHODOLOGY

This section details the proposed attentive time-aware user identity linkage across social media. Essentially, although the distributed UGCs of one user on different social media platforms may reflect different aspects of the user, they should share certain underlying patterns including writing styles, posting behaviors and personal interests, and hence coherently characterize the same user. Towards this end, we first learn the accurate representations for heterogeneous UGCs on different social media platforms. We then employ the attentive time-aware user similarity modeling to identify the same user, which jointly compiles the temporal post correlation and attentive user similarity modeling.

A. Representation Learning for UGCs

Undoubtedly, the representation learning for each UGC plays a pivotal role in user characterization. In this work, the heterogeneous UGCs on two social media platforms involve two modalities: the image and text, where UGCs on O_1 only possess the textual modality and those on O_2 include both. We thus here first learn the textural and visual representation for each modality, respectively.

Textural Representation Learning. To capture the underlying message thoroughly and characterize the user accurately, we adopt the BiLSTM framework to encode the textural modality of one's historical posts, which have achieved great success in various natural language processing tasks [49–51]. Taking the text on O_1 as an example, as that on O_2 can be derived in the same manner.

For simplicity, we temporally omit the superscript i and subscript k of t_k^i, which refers to the k-th post of user u_1^i on O_1. Given a textual post t consisting of M words, $t = \{x^1, x^2, \cdots, x^M\}$, we first embed word $x^z, z = \{1, 2, \cdots, M\}$, into the vector $\mathbf{e}^z \in \mathbb{R}^{D_e}$ with the help of the Global Vectors for Word Representation (GloVe), which is able to surpass the alternative state-of-the-art methods in the word analogy task [52]. Based on the word embeddings, we then employ the BiLSTM to learn the representation for the user post t. The key advantage of BiLSTM is to take both the forward and backward word sequence into account, which enables us to comprehensively capture the post context. Let $\overrightarrow{\mathbf{f}}^z$ and $\overleftarrow{\mathbf{f}}^z$ be the z-th hidden states of the forward LSTM and backward LSTM, respectively. Here, we take the derivation of $\overrightarrow{\mathbf{f}}^z$ as an example, as that of $\overleftarrow{\mathbf{f}}^z$ can be obtained in the similar manner. In particular, we calculate $\overrightarrow{\mathbf{f}}^z$ as follows,

$$\begin{cases} \mathbf{u}_z = \sigma(\mathbf{W}_u[\overrightarrow{\mathbf{f}}^{z-1}, \mathbf{e}^z] + \mathbf{b}_u), \\ \mathbf{r}_z = \sigma(\mathbf{W}_r[\overrightarrow{\mathbf{f}}^{z-1}, \mathbf{e}^z] + \mathbf{b}_r), \\ \mathbf{m}_z = tanh(\mathbf{W}_m[\mathbf{r}_z \odot \overrightarrow{\mathbf{f}}^{z-1}, \mathbf{e}^z] + \mathbf{b}_m), \\ \overrightarrow{\mathbf{f}}^z = \mathbf{u}_z \odot \mathbf{m}_z + (\mathbf{1} - \mathbf{u}_z) \odot \overrightarrow{\mathbf{f}}^{z-1}, \end{cases} \tag{1}$$

where $\mathbf{u}_z$ and $\mathbf{r}_z$ symbolize the update gate and reset gate, and $\mathbf{m}_z$ represents the status of the memory cell. $\mathbf{W}_u$, $\mathbf{W}_r$ and $\mathbf{W}_m$ are the weight matrices, while $\mathbf{b}_u$, $\mathbf{b}_r$ and $\mathbf{b}_m$ are the bias vectors. $\odot$ denotes the element-wise multiplication operation, $\sigma(\cdot)$ means the sigmoid activation function, and then by concatenating $\overrightarrow{\mathbf{f}}^z$ and $\overleftarrow{\mathbf{f}}^z$, we can obtain the final latent representation $\mathbf{f}^z$ for the word x^z as follows,

$$\mathbf{f}^z = [\overrightarrow{\mathbf{f}}^z, \overleftarrow{\mathbf{f}}^z]. \tag{2}$$

Ultimately, considering the limited length of the user post on social media platforms, we simply represent the user post t with the sum of the word representations as follows,

$$\tilde{\mathbf{t}} = \sum_{z=1}^{M} \mathbf{f}^z. \tag{3}$$

Accordingly, we can obtain the latent representation $\tilde{\mathbf{t}}_k^i$ for the k-th textual post t_k^i of user u_1^i and $\tilde{\mathbf{c}}_g^i$ for the textual description c_g^i generated by u_2^i.

Visual Representation Learning. As for the visual modality (i.e., image) of a given user post, we first obtain its representation with the help of Residual Network (ResNet) [53], which has delivered the superior performance in several challenging computer vision tasks on ImageNet [54] and Microsoft Common Objects in Context [55]. It is worth noting that only the multi-modal UGCs on O_2 involve images, and we thus only conduct the visual representation learning for them. In particular, we first feed the g-th image post v_g^i of user u_2^i

into the ResNet h. Hereafter, we employ a fully connected layer to project the learned visual representation into the latent common space with the textual representation. We have,

$$\tilde{\mathbf{v}}_g^i = \mathbf{W}_h h(v_g^i|\mathbf{\Theta}_r) + \mathbf{b_h}, \tag{4}$$

where $\mathbf{W}_h$ and $\mathbf{b_h}$ are the weight matrix and bias vector of the fully connected layer, and $\mathbf{\Theta}_r$ denotes the network parameter of h. Accordingly, we can acquire the latent representation $\tilde{\mathbf{v}}_g^i$ for the visual description v_g^i of the user u_2^i.

B. Attentive Time-aware User Identity Linkage

Having obtained the representations of posts on different social media platforms, we can conduct the user similarity modeling, and afterward identify if u_1^i and u_2^i refer to the same user. For simplicity, we temporally omit the superscript i in all notations.

1) One Naive Approach: One naive approach to measuring the user account similarity is to calculate the average similarity of all pair-wise user posts, i.e., t_k and (v_g, c_g) of user accounts u_1 and u_2, respectively. Adopting the widely-used cosine similarity, similar to [56, 57], we can acquire the pair-wise similarity $m_{k,g}$ between posts t_k and (v_g, c_g) as follows,

$$\begin{cases} m_{k,g} &= \frac{1}{2}m_{k,g}^c + \frac{1}{2}m_{k,g}^v, \\ m_{k,g}^c &= cos(\tilde{\mathbf{t}}_k, \tilde{\mathbf{c}}_g), \\ m_{k,g}^v &= cos(\tilde{\mathbf{t}}_k, \tilde{\mathbf{v}}_g), \end{cases} \tag{5}$$

where $m_{k,g}^c$ and $m_{k,g}^v$ correspond to the textual-textual and textual-visual similarities between posts t_k and s_g, respectively. $cos(\cdot,\cdot)$ refers to the cosine function.

By averaging all the $m_{k,g}$'s for user accounts u_1 and u_2, we can gain the user similarity. However, this method suffers from the following three limitations. 1) It treats all users' historical posts equally but overlooks the temporal correlation among their distributed posts on different social media platforms. 2) Taking different modalities (i.e., text and image) of a UGC uniformly, it neglects the confidence varies over different modalities on user characterization. And 3) it tends to learn the local pair-wise post similarity, which ignores the global user context and may lead to the degraded performance.

2) Time-aware Post Correlation Modeling: As a matter of fact, user's posts generated on different social media platforms at an adjacent period tend to be similar [16]. For example, a man may post the text "I am coming! France" on Twitter before his journey, while sharing pictures on Instagram during the trip.

Towards this end, we propose to learn the user similarity by incorporating the temporal factors, where we argue the similarity of posts generated on different social media platforms at a similar period merits our special attention. Accordingly, we introduce the time decay factor $r_{k,g}$ to capture the temporal correlation between posts t_k and s_g, which intuitively reflects the influence of their post similarity towards the final user similarity modeling between u_1 and u_2. In particular, we measure $r_{k,g}$ with the following time decay function,

$$r_{k,g} = \frac{1}{\log|p_k - q_g|}, \tag{6}$$

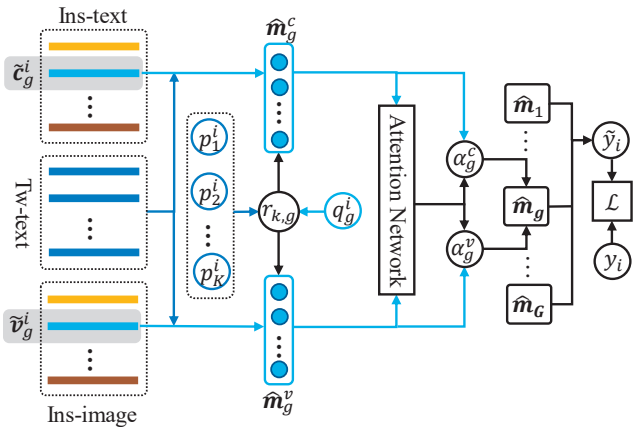

Fig. 4. Workflow of the proposed UserNet scheme, which consists of two key components: the representation learning for UGCs and the attentive time-aware user identity linkage.

where p_k and q_g are the timestamps of posts t_k and s_g, respectively. As can be seen, the smaller the timestamp difference between two posts is, the higher temporal post correlation we impose on them. This also corresponds with the larger confidence we assign towards the user similarity modeling. Thereafter, by multiplying $m_{k,g}^c$ and $m_{k,g}^v$, we can obtain the time-aware pair-wise similarities $\widehat{m}_{k,g}^c = r_{k,g}m_{k,g}^c$ and $\widehat{m}_{k,g}^v = r_{k,g}m_{k,g}^v$, respectively.

3) Attentive User Similarity Modeling: As aforementioned, each multi-modal UGC on social media O_2 involves both the image and corresponding textual description. Undoubtedly, both modalities can convey important cues regarding the user characterization. For example, the image may objectively reflect the user's current state, while the textual description can additionally reveal the subjective feeling and even the writing style of the user. Apparently, different modalities may capture users' characteristics with different confidence levels. Towards this end, to adaptively characterize the confidence of different modalities towards the user identity linkage, we propose to attentively measure the user similarity by introducing the attention mechanism to our scheme.

The attention mechanism has been proven effective in many machine learning tasks [58–64], including multimedia recommendation [62] and image caption [63, 64]. To comprehensively model the confidence of each modality regarding the user similarity modeling and boost the performance, we resort to the global post similarity distribution rather than the local pair-wise post one. As such, for each multi-modal post s_g of user account u_2, we represent its global visual similarity distribution to user account u_1 with the latent vector $\widehat{\mathbf{m}}_g^v = [\widehat{m}_{1,g}^v, \widehat{m}_{2,g}^v, \cdots, \widehat{m}_{K,g}^v]$ and that textural one with $\widehat{\mathbf{m}}_g^c = [\widehat{m}_{1,g}^c, \widehat{m}_{2,g}^c, \cdots, \widehat{m}_{K,g}^c]$, respectively. For simplicity, we further temporally omit the subscript g of $\widehat{\mathbf{m}}_g^v$ and $\widehat{\mathbf{m}}_g^c$. Given the above global visual and textural similarity of the multi-modal post, we are able to estimate the confidence of different modalities as follows,

$$\begin{cases} \mathbf{h}_v = tanh(\mathbf{W}_v\widehat{\mathbf{m}}^v + \mathbf{b}_v), \\ \mathbf{h}_c = tanh(\mathbf{W}_c\widehat{\mathbf{m}}^c + \mathbf{b}_c), \\ [\alpha_v, \alpha_c] = softmax(\mathbf{a}^{\mathrm{T}} con(\mathbf{h}_v, \mathbf{h}_c)), \end{cases} \tag{7}$$

Algorithm 1 Attentive Time-aware User Modeling.

Input: Training user account pairs $\mathcal{U}$
Output: Predicted probability $\tilde{y}_i$

Initialize neural network parameters $\mathbf{\Theta}$.
repeat:
  Draw (u_1^i, u_2^i) from $\mathcal{U}$.
  for $g = 1$ to G **do**
    for $k = 1$ to K **do**
      Compute $\widehat{m}_{k,g}^v$, $\widehat{m}_{k,g}^c$ according to Eqns. (5) and (6).
    Conduct similarity modeling according to Eqn. (8).
  Compute $\tilde{y}_i$ according to Eqn. (10).
until : Objective value converges.

where $\mathbf{W}_v$ and $\mathbf{W}_c$ are weight matrices, while $\mathbf{b}_v$ and $\mathbf{b}_c$ are bias vectors for the attention network. $con(\cdot)$ stands for the concentration operation. α_v and α_c refer to the normalized confidence of the visual and textual modalities, respectively. Here, the to-be-learned vector $\mathbf{a}$ can be interpreted as the representation of the query "*which modality of the given post conveys more significant cues towards the user identity linkage*". Ultimately, the global similarity distribution $\widehat{\mathbf{m}}$ of u_2's post s pertaining to user account u_1 is computed as follows,

$$\widehat{\mathbf{m}} = \alpha_v \widehat{\mathbf{m}}^v + \alpha_c \widehat{\mathbf{m}}^c. \tag{8}$$

Accordingly, for each multi-modal post s_g of user account u_2, we can acquire its global similarity distribution $\widehat{\mathbf{m}}_g$ towards user account u_1 and the confidences of different modalities (i.e., α_g^c and α_g^v).

By now, we can proceed to the final user similarity modeling between user accounts u_1 and u_2. In particular, we integrate the global similarity distributions of all posts belonging to u_2 by $\mathbf{d} = [d_1, d_2, \cdots, d_G] \in \mathbb{R}^G$, where $d_g = avg(\widehat{\mathbf{m}}_g)$ refers to the average pooling over the distribution vector $\widehat{\mathbf{m}}_g$, $g = \{1, 2, \cdots, G\}$. Thereinto, K and G refer to the total number of posts of u_1 and u_2, respectively. In a sense, $\mathbf{d}$ can be treated as the latent similarity representation between u_1 and u_2. Thereafter, we utilize the multi-layer perceptron (MLP) [65] to project the latent similarity space into the probability space as follows,

$$\tilde{y} = sigmoid(\mathbf{w}^T \mathbf{d} + b), \tag{9}$$

where $\tilde{y}$ is the predicted probability that user accounts u_1 and u_2 refer to the same user identity. Meanwhile, $\mathbf{w}$ and b are the weight vector and bias, respectively. Ultimately, on the basis of the cross entropy loss [66], we reach the final objective function for binary classification,

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} [-y_i \log \tilde{y}_i - (1 - y_i) \log(1 - \tilde{y}_i)]. \tag{10}$$

Figure 4 illustrates the workflow of our scheme, while the procedure of attentive time-aware user identity linkage scheme is summarized in Algorithm 1.

V. Experiment

In this section, we conducted extensive experiments to answer the following three research questions:

- **RQ1.** Does our UserNet outperform the state-of-the-art methods?
- **RQ2.** How does the attentive user similarity modeling affect the performance of UserNet?
- **RQ3.** How do temporal factors influence the performance of UserNet?

A. Dataset

As a matter of fact, various user identity linkage datasets have been collected for different research purposes, including Bennacer [67], Zhang [68] and Yan [3]. However, the currently released datasets mainly focus on collecting users' profiles or network structures, leaving out the plentiful UGCs. Although Yan [2] has tried to conduct the dataset based on UGCs, they ignored the timestamps. Towards this end, to facilitate the research line of user identity linkage from UGCs, we created a dataset, named TWIN, from two popular heterogeneous social media platforms: Twitter and Instagram. In particular, we first obtained the user account mappings on Twitter and Instagram from "#mytweet via Instagram" [69], which comprises 15,595 users with their accounts on six social media platforms. For each mapping pair, we collected the user's latest 200 UGCs with timestamps on both Twitter and Instagram within the period from 17/03/2009 to 07/11/2018. To ensure the quality of our dataset, we only retained users that have more than 150 UGCs on both social media platforms. Ultimately, we obtained 5,765 user pairs with 1,729,500 UGCs and corresponding timestamps on Twitter and Instagram. In particular, we noticed that words are flexible and variant in UGCs, which causes great difficulties in user characterization. Therefore, we adopted TextBlob [70] to unify words into lowercase letters, and removed emoji in UGCs.

B. Experiment Settings

Visual Feature. To better characterize an image, we turned to the ResNet, which is efficient in multiple computer vision challenges, such as ImageNet [54]. In particular, we adopted the pre-trained ResNet with 152 layers. Each image of the UGC, resized to 244×244, is fed into the ResNet and embedded as a 2,048-D vector.

Textual Feature. To generate the textual embeddings for UGCs, we employed the GloVe with 100-D vector, pre-trained on Twitter. For words unavailable in the vocabulary of GloVe, we simply represented them with all-zeros placeholder vectors.

Optimization. We divided our user account pairs into three chunks: 80% for training, 10% for validation and 10% for testing. These are treated as the positive pairs, while the negative unmatched pairs are generated randomly. The amount of positive and negative pairs is the same. As we cast the user identity linkages as a binary classification, we adopted the accuracy as the evaluation metric. For optimization, we employed the Adam Optimizer with the learning rate being 0.0001. We utilized the idea of grid search to determine the

TABLE II
PERFORMANCE COMPARISON AMONG DIFFERENT MODELS IN TERMS OF ACCURACY.

Model	Accuracy
QDA	0.5625
WHOLE	0.6836
WSF-GBDT	0.7552
BPR-DAE	0.5703
UserNet	**0.8369**

TABLE III
PERFORMANCE OF USERNET AND ITS DERIVATIVES WITH DIFFERENT NUMBERS OF POSTS UTILIZED.

Model	K=G=150	K=G=100	K=G=50
UserNet	**0.8369**	**0.8047**	**0.7939**
UserNet-NoA	0.8281	0.7881	0.7832
UserNet-NoC	0.5615	0.5576	0.5703
UserNet-NoV	0.8223	0.7968	0.7900

optimal values for hyperparameters including the batch size, the learning rate and the number of hidden units. The proposed scheme is fine-tuned on the training set and validation set, and the performance on the testing set is reported. All the experiments are mainly conducted over a server equipped with 32 Cores Intel (R) Xeon (R) CPU E5-2620 v4 2.10GHz and four TITAN Xp GPUs. In particular, our model has around 8.4M parameters, and the running time of identifying whether a given user account pair refers to the same identity based on their pre-extracted visual and textual features is 25.32ms. Hence, our model is practical for real-world scenarios with reasonable time efficiency.

C. On Model Comparison (RQ1)

As limited efforts worked on the problem of user identity linkage from UGCs, we compared our scheme (UserNet) with the following baselines.

- **QDA.** The first baseline treats all UGCs of one user account as a whole, where the textual posts are merged together and embedded by Doc2Vec [71], and the image representations of one user extracted by ResNet are concatenated and then projected to the low dimensional space by the principal component analysis [72]. Accordingly, we derived the user representation on O_2 by fusing his/her multi-modal representations on O_2, and then obtained the user similarity by quadratic discriminant analysis [73]. To be specific, we programmed it with the help of scikit-learn [74].
- **WHOLE.** Similar to QDA, this baseline also characterizes each user account's textual/visual modality by considering all UGCs as a whole. Different from QDA, WHOLE employs BiLSTM to derive the textual representation rather than Doc2Vec in QDA. Besides, WHOLE conducts the text-text similarity and the text-image similarity separately by MLP, and obtains the final user similarity by fusing the text-text similarity and the text-image similarity.
- **WSF-GBDT.** The third baseline is derived from the method in [75], which introduced four writing-style features, including lexical, syntactic, structural, and content-specific features, to characterize users and employed the support vector machine (SVM) [76] for the user identity classification. We altered the shallow learning method with the gradient boosting classifier [77], which surpasses SVM in our experiments.
- **BPR-DAE.** We chose the content-based neural method proposed by [78] that explores the heterogeneous data for the pair-wise item modeling based on the Bayesian Personalized Ranking (BPR) framework [79]. In particular, we adapted the scheme to user identity linkage by calculating the average similarity of all pair-wise user posts. Essentially, the post similarity takes different modalities (i.e, text and image) uniformly. Different from our scheme, this baseline ignores the temporal post correlation and attentive user similarities.

Table II shows the performance comparison of different methods. Based on that table, we had the following observations: 1) Both relying on the shallow machine learning technique, WSF-GBDT performs much better than QDA, which suggests the superiority of the hand-crafted writing-style features over the representations derived from the pre-trained Doc2Vec. 2) UserNet significantly outperforms the best baseline WSF-GBDT. This may be attributed to the fact that UserNet integrates the multi-modal UGCs rather than the only writing-style features. 3) UserNet surpasses WHOLE, indicating the advantage of the fine-grained post similarity modeling in the context of user identity linkage, which can further incorporate the temporal post correlation. And 4) UserNet shows better performance than BPR-DAE that only focuses on the local pair-wise post similarity modeling. This result reflects the effectiveness of global post-user similarity modeling towards the user identity linkage.

D. On Attention Mechanism (RQ2)

To evaluate the effectiveness of our attentive user similarity modeling, we compared UserNet with the following three derivations. 1) **UserNet-NoA**. We disabled the attention mechanism by uniformly assigning the confidence of different modalities in user similarity modeling. 2) **UserNet-NoC**. We made our UserNet simply focus on the post similarity distribution derived from the visual modality by discarding that from the textual modality of UGCs on O_2. 3) **UserNet-NoV**. Similar to UserNet-NoC, we adapted UserNet to only pay attention to the similarity distribution derived from the textual modality.

Table III shows the performance of UserNet and its above derivatives with different numbers of posts utilized. Firstly, as can be seen, the proposed UserNet consistently shows superiority over UserNet-NoA with different numbers of posts, which enables us to safely draw the conclusion that it does exist different confidence over the image and the textual

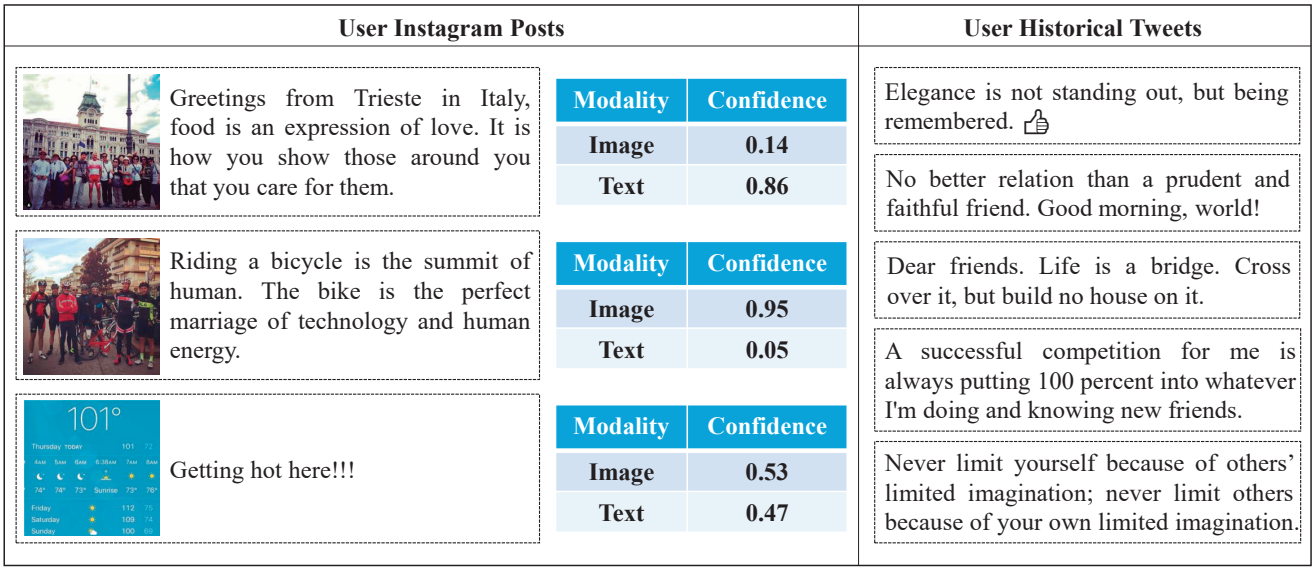

Fig. 5. Illustration of different confidence over different modalities of the user's Instagram posts.

description in a user post pertaining to the user characterization. Hence, it is essential to attentively combine the post similarity derived from different modalities rather than uniformly. Secondly, we found that UserNet outperforms both UserNet-NoC and UserNet-NoV consistently, which suggests that the visual and textual information complements each other and both contribute to the user characterization when the temporal post correlation is taken into consideration. Interestingly, we observed that UserNet-NoV outperforms UserNet-NoC, indicating that the textual modality of UGCs on O_2 contributes more on the user similarity modeling, as compared with the visual modality. The possible reasons are twofold. 1) It is easier to capture the post similarity based on the homogeneous data (i.e., Twitter text and Instagram text) than that on heterogeneous data (i.e., Twitter text and Instagram image). And 2) the textual information of one UGC is more concise and descriptive to present the user characterization, such as the location, occupation and even the relationship. Thirdly, we found that the more posts we utilized, the higher performance can be achieved by almost all methods, except UserNet-NoC where the textual modality is removed. This reflects the potential of UGCs in the user characterization and verifies the dominant role of the textual modality in user identity linkage again. Last but not least, we noticed that even only with the textual modality, UserNet-NoV can achieve better performance than UserNet-NoA. This indicates that uniformly fusing the post similarities from both modalities may hurt the performance that simply based on the more powerful modality (i.e., text in our case).

To gain a better understanding of the attention mechanism in our scheme, we provided the experimental results on modality confidence assignment with UGCs of one user in Figure 5. To be specific, we can extract the confidences of different modalities according to Eqn. (7) on the basis of the TensorFlow framework. Due to the limited space, we only listed several historical tweets of the user as examples. As can be seen, different levels of confidence can be assigned to different modalities of a UGC. In particular, for the first Instagram post of the user, we noticed that higher confidence is assigned to its textual description as compared to its image. Checking the user's historical tweets, we found that this user seems to be a positive guy, which characteristic can be reflected better by the greeting words than the image of the Instagram post. Therefore, it is reasonable to attentively assign higher confidence to the textual modality. Similar observations can be found in other examples.

TABLE IV
PERFORMANCE COMPARISON OF OUR SCHEME WITH ITS VARIANTS EXCLUDING THE TIME FACTORS.

Model	Visual	Text	All
UserNet	**0.5615**	**0.8223**	**0.8369**
UserNet-NoT	0.5166	0.8193	0.7217

E. On Temporal Post Correlation (RQ3)

To verify the effectiveness of the time factors, we also conducted experiments over the temporal correlation comparison. In particular, we adopted the baseline UserNet-NoT, where we disabled the time factors in UserNet by removing the time decay parameters. Table IV illustrates the effects of the temporal correlation in our scheme with different modality configurations. As can be seen, our scheme consistently shows superiority over UserNet-NoT across different modality configurations, demonstrating that the temporal post correlation plays a pivotal role in the final user similarity modeling. Interestingly, we found that the UserNet-NoT only with the textual posts outperforms that with all posts, which suggests that the visual signal is not much reliable for user identity linkage and the general heterogeneous data fusion cannot

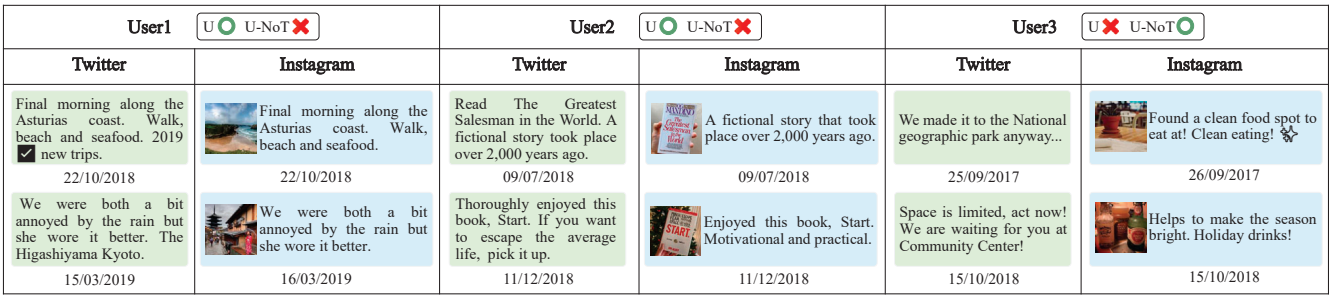

Fig. 6. Illustration of the comparison between UserNet and UserNet-NoT on testing pairs. "U" and "U-NoT" refer to the UserNet and UserNet-NoT, respectively. We represent the correct judgments of the model with the green circle and that of the wrong with the red cross.

guarantee the performance boost. It also reflects the necessity of incorporating the time factor in the post similarity modeling.

To gain more deep insights on the influence of time factors, we illustrated the comparison between UserNet and UserNet-NoT on several testing user account pairs in Figure 6. As we can see, UserNet performs much better than UserNet-NoT in cases when the user posted similar UGCs across different social media during the same period. For example, we found that the user1 recorded his/her final morning at Asturias coast on the same day on both Twitter and Instagram with almost exact textual descriptions. Similar observation can be found in his/her later post regarding the annoying rain. In addition, we have the similar observation for the user2 in Figure 6, which validates the advantage of taking the temporal post correlation in the user similarity modeling. Nevertheless, the temporal post correlation may also lead to the failed cases, such as the user3 in Figure 6, who usually posted totally different contents on different social media platforms in the same period.

VI. Conclusion and Future Work

In this paper, we studied the problem of user identity linkage across social media based on the heterogeneous UGCs. In particular, we proposed an attentive time-aware user identity linkage scheme, which jointly unifies the time-aware post correlation modeling and attentive user similarity modeling. Towards the evaluation, we created a large-scale user identity linkage dataset, TWIN, comprising 5,765 users with 1,729,500 UGCs and corresponding timestamps from Twitter and Instagram. Extensive experiments have been carried out and demonstrated the effectiveness of our proposed scheme. Interestingly, we found that the visual and textual information of a post complement each other and both contribute to the user characterization. In addition, different modalities in a user post do have different confidence in user similarity modeling. Meanwhile, the experimental results show that the temporal post correlation plays a pivotal role in user similarity modeling. As a residual product, we have released the dataset, codes, and parameters to facilitate other researchers.

Currently, we focus on tackling the user identity linkage problem based on UGCs. In this work, we mainly investigate the potential of users' distributed heterogeneous contents on different platforms in the user identity linkage problem, especially for active users. In the future, we plan to further incorporate the users' profile and social network structure to take users with sparse UGCs into consideration.

References

[1] S. Kumar, R. Zafarani, and H. Liu, "Understanding user migration patterns in social media," in *Proceedings of the Conference on Artificial Intelligence*. AAAI Press, 2011, pp. 1204–1209.

[2] M. Yan, J. Sang, and C. Xu, "Unified youtube video recommendation via cross-network collaboration," in *Proceedings of the ACM International Conference on Multimedia Retrieval*. ACM, 2015, pp. 19–26.

[3] J. S. Ming Yan and C. Xu, "Mining cross-network association for youtube video promotion," in *Proceedings of the ACM International Conference on Multimedia*. ACM, 2014, pp. 557–566.

[4] J. Sang, Z. Deng, D. Lu, and C. Xu, "Cross-osn user modeling by homogeneous behavior quantification and local social regularization," *IEEE Transactions on Multimedia*, vol. 17, no. 12, pp. 2259–2270, 2015.

[5] J. Sang, M. Yan, and C. Xu, "Understanding dynamic cross-osn associations for cold-start recommendation," *IEEE Transactions on Multimedia*, vol. 20, no. 12, pp. 3439–3451, 2018.

[6] M. Yan, J. Sang, C. Xu, and M. S. Hossain, "A unified video recommendation by cross-network user modeling," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 12, no. 4, pp. 53:1–53:24, 2016.

[7] M. Yan, J. Sang, C. Xu, and M. S. Hossain, "Youtube video promotion by cross-network association: @britney to advertise gangnam style," *IEEE Transactions on Multimedia*, vol. 17, no. 8, pp. 1248–1261, 2015.

[8] D. Perito, C. Castelluccia, M. A. Kaafar, and P. Manils, "How unique and traceable are usernames?" pp. 1–17, 2011.

[9] X. Zhou, L. Xun, H. Zhang, and Y. Ma, "Cross-platform identification of anonymous identical users in multiple social media networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 2, pp. 411–424, 2016.

[10] Z. Rong, J. Li, H. Chen, and H. Zan, "A framework for authorship identification of online messages: Writing-style features and classification techniques," *Journal of the Association for Information Science and Technology*, vol. 57, no. 3, pp. 378–393, 2010.

[11] M. Tong, H. Shen, S. Liu, X. Jin, and X. Cheng, "Predict anchor links across social networks via an embedding approach," in *Proceedings of the International Joint Conference on Artificial Intelligence*. IJCAI/AAAI Press, 2016, pp. 1823–1829.

[12] X. Mu, F. Zhu, E. P. Lim, J. Xiao, and Z. H. Zhou, "User identity linkage by latent user space modelling," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 1775–1784.

[13] S. Ahangama and D. C. C. Poo, "Cross domain approximate matching of user shared content for user entity resolution,"

in *Proceedings of the International Conference on Information Systems*. Association for Information Systems, 2018.

[14] P. Liu, X. Qiu, and X. Huang, "Recurrent neural network for text classification with multi-task learning," in *Proceedings of the International Joint Conference on Artificial Intelligence*. AAAI Press, 2016, pp. 2873–2879.

[15] J. I. Olszewska, "Designing transparent and autonomous intelligent vision systems," in *Proceedings of the International Conference on Agents and Artificial Intelligence*. SciTePress, 2019, pp. 850–856.

[16] S. Liu, S. Wang, F. Zhu, J. Zhang, and R. Krishnan, "HYDRA: large-scale social identity linkage via heterogeneous behavior modeling," in *Proceedings of the International Conference on Management of Data*. ACM, 2014, pp. 51–62.

[17] Y. Chen, C. Zhuang, Q. Cao, and P. Hui, "Understanding cross-site linking in online social networks," in *Proceedings of the Workshop on Social Network Mining and Analysis*. ACM, 2014, pp. 6:1–6:9.

[18] O. Goga, H. Lei, S. H. K. Parthasarathi, G. Friedland, R. Sommer, and R. Teixeira, "Exploiting innocuous activity for correlating users across sites," in *Proceedings of the International World Wide Web Conference*. ACM, 2013, pp. 447–458.

[19] I. P. Fellegi and A. B. Sunter, "A theory for record linkage," *Journal of the American Statistical Association*, vol. 64, pp. 1183–1210, 1969.

[20] R. Zafarani and H. Liu, "Connecting corresponding identities across communities," in *Proceedings of the International Conference on Weblogs and Social Media*. The AAAI Press, 2009.

[21] K. Shu, S. Wang, J. Tang, R. Zafarani, and H. Liu, "User identity linkage across online social networks: A review," *SIGKDD Explorations*, vol. 18, no. 2, pp. 5–17, 2016.

[22] T. Alqaisi, D. Gledhill, and J. I. Olszewska, "Embedded double matching of local descriptors for a fast automatic recognition of real-world objects," in *Proceedings of the IEEE International Conference on Image Processing*. IEEE, 2012, pp. 2385–2388.

[23] C. Faloutsos, R. Barber, M. Flickner, J. Hafner, W. Niblack, D. Petkovic, and W. Equitz, "Efficient and effective querying by image content," *Journal of Intelligent Information Systems*, vol. 3, no. 3/4, pp. 231–262, 1994.

[24] A. G. Bors and A. Papushoy, "Image retrieval based on query by saliency content," in *Visual Content Indexing and Retrieval with Psycho-Visual Models*. Springer, 2017, pp. 171–209.

[25] C. E. Jacobs, A. Finkelstein, and D. Salesin, "Fast multiresolution image querying," in *Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques*. ACM, 1995, pp. 277–286.

[26] E. J. Guglielmo and N. C. Rowe, "Natural-language retrieval of images based on descriptive captions," *ACM Transactions on Information Systems*, vol. 14, no. 3, pp. 237–267, 1996.

[27] J. I. Olszewska and D. Wilson, "Hausdorff-distance enhanced matching of scale invariant feature transform descriptors in context of image querying," in *Proceedings of the IEEE International Conference on Intelligent Engineering Systems*. IEEE, 2012, pp. 91–96.

[28] X. Cao and Y. Yu, "BASS: A bootstrapping approach for aligning heterogeneous social networks," in *Machine Learning and Knowledge Discovery in Databases - European Conference*. Springer, 2016, pp. 459–475.

[29] X. Kong, J. Zhang, and P. S. Yu, "Inferring anchor links across multiple heterogeneous social networks," in *Proceedings of the ACM International Conference on Information and Knowledge Management*. ACM, 2013, pp. 179–188.

[30] J. Zhang and P. S. Yu, "Integrated anchor and social link predictions across social networks," in *Proceedings of the International Joint Conference on Artificial Intelligence*. AAAI Press, 2015, pp. 2125–2132.

[31] Y. Ren, C. Aggarwal, and J. Zhang, "Activeiter: Meta diagram based active learning in social networks alignment," *IEEE Transactions on Knowledge and Data Engineering*, 2019.

[32] C. Li, S. Wang, P. S. Yu, L. Zheng, X. Zhang, Z. Li, and Y. Liang, "Distribution distance minimization for unsupervised user identity linkage," in *Proceedings of the ACM International Conference on Information and Knowledge Management*. ACM, 2018, pp. 447–456.

[33] W. Wang, H. Yin, X. Du, W. Hua, Y. Li, and Q. V. H. Nguyen, "Online user representation learning across heterogeneous social networks," in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2019, pp. 545–554.

[34] P. Cui, S. Liu, and W. Zhu, "General knowledge embedded image representation learning," *IEEE Transactions on Multimedia*, vol. 20, no. 1, pp. 198–207, 2018.

[35] S. Wang and W. Guo, "Sparse multigraph embedding for multimodal feature representation," *IEEE Transactions on Multimedia*, vol. 19, no. 7, pp. 1454–1466, 2017.

[36] Y. He, S. Xiang, C. Kang, J. Wang, and C. Pan, "Cross-modal retrieval via deep and bidirectional representation learning," *IEEE Transactions on Multimedia*, vol. 18, no. 7, pp. 1363–1377, 2016.

[37] F. Wu, X. Lu, J. Song, S. Yan, Z. M. Zhang, Y. Rui, and Y. Zhuang, "Learning of multimodal representations with random walks on the click graph," *IEEE Transactions on Multimedia*, vol. 25, no. 2, pp. 630–642, 2016.

[38] C. Hsu and C. Lin, "Cnn-based joint clustering and representation learning with feature drift compensation for large-scale image data," *IEEE Transactions on Multimedia*, vol. 20, no. 2, pp. 421–429, 2018.

[39] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural Networks*, vol. 18, no. 5-6, pp. 602–610, 2005.

[40] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the Annual Conference on Neural Information Processing Systems*. Curran Associates, Inc., 2012, pp. 1097–1105.

[41] K. Georgiev and P. Nakov, "A non-iid framework for collaborative filtering with restricted boltzmann machines," in *Proceedings of the International Conference on Machine Learning*. JMLR.org, 2013, pp. 1148–1156.

[42] D. Wang, C. Peng, and W. Zhu, "Structural deep network embedding," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 1225–1234.

[43] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the International Conference on Machine Learning*. Omnipress, 2011, pp. 689–696.

[44] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, 2014, pp. 1725–1732.

[45] A.-A. Liu, W.-Z. Nie, Y. Gao, and Y.-T. Su, "Multi-modal clique-graph matching for view-based 3d model retrieval," *IEEE Transactions on Image Processing*, vol. 25, no. 5, pp. 2103–2116, 2016.

[46] J. Rajendran, M. M. Khapra, S. Chandar, and B. Ravindran, "Bridge correlational neural networks for multilingual multimodal representation learning," pp. 171–181, 2016.

[47] D. Wang, P. Cui, M. Ou, and W. Zhu, "Deep multimodal hashing with orthogonal regularization," in *Proceedings of the International Joint Conference on Artificial Intelligence*. AAAI Press, 2015, pp. 2291–2297.

[48] J. Zhou and J. Fan, "Translink: User identity linkage across heterogeneous social networks via translating embeddings," in *Proceedings of the IEEE Conference on Computer Communications*. IEEE, 2019, pp. 2116–2124.

[49] Y. Yao and H. Zheng, "Bi-directional lstm recurrent neural network for chinese word segmentation," in *Proceedings of the International Conference on Neural Information Processing*. Springer, 2016, pp. 345–353.
[50] A. R. Mohamed, F. Seide, Y. Dong, J. Droppo, A. Stoicke, G. Zweig, and G. Penn, "Deep bi-directional recurrent networks over spectral windows," in *IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2016, pp. 78–83.
[51] Q. Wang, T. Luo, W. Dong, and X. Chao, "Chinese song iambics generation with neural attention-based model," in *Proceedings of the International Joint Conference on Artificial Intelligence*. IJCAI/AAAI Press, 2016, pp. 2943–2949.
[52] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. ACL, 2014, pp. 1532–1543.
[53] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, 2016, pp. 770–778.
[54] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.
[55] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proceedings of the European Conference on Computer Vision*. Springer, 2014, pp. 740–755.
[56] M. Carvalho, R. Cadène, D. Picard, L. Soulier, N. Thome, and M. Cord, "Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings," pp. 35–44, 2018.
[57] M. Liu, X. Wang, L. Nie, X. He, B. Chen, and T.-S. Chua, "Attentive moment retrieval in videos," in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2018, pp. 15–24.
[58] K. Cho, A. C. Courville, and Y. Bengio, "Describing multimedia content using attention-based encoder-decoder networks," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 1875–1886, 2015.
[59] Y. Jiao, Z. Li, S. Huang, X. Yang, B. Liu, and T. Zhang, "Three-dimensional attention-based deep ranking model for video highlight detection," *IEEE Transactions on Multimedia*, vol. 20, no. 10, pp. 2693–2705, 2018.
[60] B. Zhao, X. Wu, J. Feng, Q. Peng, and S. Yan, "Diversified visual attention networks for fine-grained object classification," *IEEE Transactions on Multimedia*, vol. 19, no. 6, pp. 1245–1256, 2017.
[61] Z. Fan, X. Zhao, T. Lin, and H. Su, "Attention-based multiview re-observation fusion network for skeletal action recognition," *IEEE Transactions on Multimedia*, vol. 21, no. 2, pp. 363–374, 2019.
[62] J. Chen, H. Zhang, X. He, L. Nie, W. Liu, and T.-S. Chua, "Attentive collaborative filtering: Multimedia recommendation with item-and component-level attention," in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2017, pp. 335–344.
[63] Z. Zhang, Q. Wu, Y. Wang, and F. Chen, "High-quality image captioning with fine-grained and semantic-guided visual attention," *IEEE Transactions on Multimedia*, vol. 21, no. 7, pp. 1681–1693, 2019.
[64] J. H. Tan, C. S. Chan, and J. H. Chuah, "COMIC: toward A compact image captioning model with attention," *IEEE Transactions on Multimedia*, vol. 21, no. 10, pp. 2686–2696, 2019.
[65] F. Rosenblatt, "Principles of neurodynamics. perceptrons and the theory of brain mechanisms," Tech. Rep., 1961.
[66] C. H. Li and C. K. Lee, "Minimum cross entropy thresholding," *Pattern Recognition*, vol. 26, no. 4, pp. 617–625, 1993.
[67] N. Bennacer, C. Nana Jipmo, A. Penta, and G. Quercini, "Matching user profiles across social networks," in *Proceedings of the International Conference on Advanced Information Systems Engineering*. Springer, 2014, pp. 424–438.
[68] Y. Zhang, J. Tang, Z. Yang, J. Pei, and P. S. Yu, "Cosnet: Connecting heterogeneous social networks with local and global consistency." in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 1485–1494.
[69] B. H. Lim, D. Lu, T. Chen, and M.-Y. Kan, "#mytweet via instagram: Exploring user behaviour across multiple social networks," in *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ACM, 2015, pp. 113–120.
[70] S. Loria, P. Keen, M. Honnibal, R. Yankovsky, D. Karesh, E. Dempsey *et al.*, "Textblob: simplified text processing," *Secondary TextBlob: Simplified Text Processing*, 2014.
[71] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proceedings of the International Conference on Machine Learning*. JMLR.org, 2014, pp. 1188–1196.
[72] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.
[73] P. A. Lachenbruch and M. Goldstein, "Discriminant analysis," *Biometrics*, vol. 35, no. 1, pp. 69–85, 1979.
[74] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2825–2830, 2011.
[75] R. Zheng, J. Li, H. Chen, and Z. Huang, "A framework for authorship identification of online messages: Writing-style features and classification techniques," *Journal of the American Society for Information Science and Technology*, vol. 57, no. 3, pp. 378–393, 2006.
[76] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
[77] F. J. G. F. APPROXIMATION, "A gradient boosting machine," *The Annals of Statistics*, 1999.
[78] X. Song, F. Feng, J. Liu, Z. Li, L. Nie, and J. Ma, "Neurostylist: Neural compatibility modeling for clothing matching," in *Proceedings of the ACM International Conference on Multimedia*. ACM, 2017, pp. 753–761.
[79] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "Bpr: Bayesian personalized ranking from implicit feedback," in *Proceedings of the Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2009, pp. 452–461.

Xiaolin Chen received the B.E. degree from Shandong University of Science and Technology in 2018. She is currently working toward the M.S. degree in the school of Computer Science and Technology, Shandong University. Her research interests include the information retrieval and social network analysis.

Xuemeng Song received the B.E. degree from University of Science and Technology of China in 2012, and the Ph.D. degree from the School of Computing, National University of Singapore in 2016. She is currently an assistant professor of Shandong University, Jinan, China. Her research interests include the information retrieval and social network analysis. She has published several papers in the top venues, such as ACM SIGIR, MM and TOIS. In addition, she has served as reviewers for many top conferences and journals.

Siwei Cui received the B.E. degree in the school of Computer Science and Technology at Shandong University in 2019. He is currently a Ph.D. student at Texas A&M University. His research interests include natural language processing, social media computing, and software engineering.

Tian Gan received her B.Sc. from East China Normal University in 2010, and the Ph.D. degree from National University of Singapore, Singapore, in 2015. She was a Research Scientist in Institute for Infocomm Research (I2R), Agency for Science, Technology and Research (A*STAR). She is currently an Assistant Professor with the School of Computer Science and Technology, Shandong University. Her research interests include multisensor event analysis, social signal processing, and wearable computing.

Zhiyong Cheng is currently a Professor with Shandong Artificial Intelligence Institute, Qilu University of Technology (Shandong Academy of Sciences). He received the Ph.D degree in computer science from Singapore Management University in 2016, and then worked as a Research Fellow in National University of Singapore. His research interests mainly focus on large-scale multimedia content analysis and retrieval. His work has been published in a set of top forums, including ACM SIGIR, MM, WWW, TOIS, IJCAI, TKDE, and TCYB. He has served as the PC member for several top conferences such as MM, MMM etc., and the regular reviewer for journals including TKDE, TIP, TMM etc.

Liqiang Nie is currently a professor with the School of Computer Science and Technology, Shandong University. Meanwhile, he is the adjunct dean with the Shandong AI institute. He received his B.Eng. and Ph.D. degree from Xi'an Jiaotong University in July 2009 and National University of Singapore (NUS) in 2013, respectively. After PhD, Dr. Nie continued his research in NUS as a research follow for more than three years. His research interests lie primarily in multimedia computing and information retrieval. Dr. Nie has co-/authored more than 160 papers, received more than 5300 Google Scholar citations as of Oct. 2019. He is an AE of Information Science, an area chair of ACM MM 2018, a special session chair of PCM 2018, a PC chair of ICIMCS 2017. Meanwhile, he is supported by the program of "Thousand Youth Talents Plan 2016", "Qilu Scholar 2016", and "The Shandong Province Science Fund for Distinguished Young Scholars 2018". In 2017, he co-founded "Qilu Intelligent Media Forum".