This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Siamese Neural Networks for User Identity Linkage Through Web Browsing

Yuanyuan Qiao, Yuewei Wu, Fan Duo, Wenhui Lin, and Jie Yang

***Abstract*— Linking online identities of users among countless heterogeneous network services on the Internet can provide an explicit digital representation of users, which can benefit both research and industry. In recent years, user identity linkage (UIL) through the Internet has become an emerging task with great potential and many challenges. Existing works mainly focus on online social networks that consider inconsistent profiles, content, and networks as features or use sparse location-based data sets to link the online behaviors of a real person. To extend the UIL problem to a general scenario, we try to link the web-browsing behaviors of users, which can help to distinguish specific users from others, such as children or malicious users. More specifically, we propose a Siamese neural network (NN) architecture-based UIL (SAUIL) model that learns and compares the highest-level feature representation of input web-browsing behaviors with deep NNs. Although the number of matching and nonmatching pairs for the UIL problem is highly imbalanced, previous studies have not considered imbalanced UIL data sets. Therefore, we further address the imbalanced learning issue by proposing cost-sensitive SAUIL (C-SAUIL) model, which assumes higher costs for misclassifying the minority class. In the experiments, the proposed model is robust and exhibits a good performance on very large, real-world data sets collected from different regions with distinct characteristics.**

***Index Terms*— Cost-sensitive classification, loss function, Siamese neural networks (NNs), user identity linkage (UIL).**

I. Introduction

THE Internet brings us ubiquitous connections with unlimited information and resources. With the prevailing use of smart devices in recent years [1], users like to carry mobile devices and connect to mobile Internet whenever and wherever possible. To provide better recommendations or services, the complete description for the user's online interests, often referred to as a user profile, is crucial. However, unlike physical trajectories, our "Digital Footprints" [2] can appear on any web service on the Internet, and the web-browsing behaviors of users are split by different web services the user have visited and various devices the user have used. The multiple states of users, including logging in, not logging in, or anonymous, make the problem of UIL more difficult. Even so, the UIL problem has attracted increasing attention in recent years [3] due to service quality (e.g., enhancing recommendations, solving cold-start problems, and web development analyses) [3], [4], research purposes (e.g., information diffusion, network dynamics, and patterns of user migration between multiple online services) [3] and cybersecurity issues (e.g., verifying ages online) [5].

Most of the previous works focus on UIL across online social networks, which consider profiles, content, and network as the components of a user's identity [3], [6]. Facing inaccurately filled profile information, heterogeneous networks, and the intuition that people's core interests will not change in short period, some researchers try to model a user's intrinsic characteristics with the user's generated or browsed content [4], [7], [8]. For the general cross-domain case, it has been found that location-based data sets are more suitable to establish users' uniqueness or classify users into different groups [9]–[11]. Toward the usage of Internet services across varied devices, the web-browsing behaviors of users are also a useful source to solve the UIL problem since they can be obtained from web logs that continuously record all Internet usage behaviors of users [12].

Although web browsing involves private information such as browsing history, UIL through web browsing has important significance for scenarios such as constructing user identity libraries to identify specific users in company or home networks, which ensures the security of the network environment for company staff and children, respectively. For example: 1) at home, family members usually share devices such as mobile phones, iPads, and computers if children visit webpages for adults on several devices without needing to log in or by using their parents' accounts; their online identities should be identified by distinguishing their browsing behaviors from those of others and 2) at the workplace, a large number of users anonymously connect to the Internet through an access point (AP) or WiFi with a dynamic Internet protocol (IP) address. When malicious online behavior is detected, it is necessary to identify all the online behaviors of suspicious users. For the above scenarios, in our study, we aim to identify

Manuscript received June 30, 2018; revised December 20, 2018, March 26, 2019 and June 16, 2019; accepted July 5, 2019. This work was supported in part by the National Natural Science Foundation of China under Grant 61701031, in part by the Funds of the Beijing Laboratory of Advanced Information Networks of the Beijing University of Posts and Telecommunications (BUPT), in part by the Funds of the Beijing Key Laboratory of Network System Architecture and Convergence of BUPT, in part by the National Key Research and Development Program of China under Grant 2017YFB0802701, in part by the Fundamental Research Funds for the Central Universities under Grant 2018XKJC04, and in part by the 111 Project of China under Grant B17007. *(Corresponding authors: Yuanyuan Qiao; Jie Yang.)*

Y. Qiao, Y. Wu, and J. Yang are with the Artificial Intelligence Institute, Beijing University of Posts and Telecommunications (BUPT), Beijing 100876, China (e-mail: yyqiao@bupt.edu.cn; janeyang@bupt.edu.cn).

F. Duo is with the Jacobs School of Engineering, University of California San Diego, San Diego, CA 92093 USA (e-mail: fduo@eng.ucsd.edu).

W. Lin is with the Technology Research Institute, Aisino Corporation, Beijing 100195, China.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2019.2929575

2162-237X © 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

target users (children or suspicious users) by distinguishing their web-browsing behaviors as accurately as possible, which will be recorded continuously for constructing complete identity libraries to identify their newly generated web-browsing behaviors under the state of anonymity or using other people's online accounts on different devices. In addition, although digital identities such as IP, spatial, and user device-related features can be extracted from web logs [13], we only focus on uniform resource locator (URL)/web content and the time the user browsed it. We try not to involve personal identifiers as the general data protection regulation (GDPR) [14] suggests.

The UIL problem is quite challenging since information generated by users online is incomplete, inconsistent, heterogeneous, and sparse; sometimes this information can even be deliberately faked. The core idea of linking a user's online identity [15] is extracting unique features that can differentiate each person, and then features belonging to the same real person are linked. Features that may show the unique nature of a person include personal profiles, generated or browsed content, social network interactions, physical locations, and timestamps. Then, the UIL problem is turned into a typical classification problem, which is usually solved by classification/prediction models. Following the idea that the intrinsic interests of humans remain unchanged over a long period [7] to interpret the deep semantics hidden in many web-browsing behaviors, in this work, we employ the Siamese neural network (NN) architecture with bidirectional long short-term memory (BLSTM). It computes the similarity between the highest-level feature representation [16]. Instead of knowing the exact class of each input, Siamese nets output the similarity value of each input pair. Siamese nets show great potential in scenarios involving finding similarities or a relationship between two comparable things, which can be applied to our UIL problem, i.e., identifying the specific user by linking all of his/her web-browsing behaviors.

Through all the research on the UIL problem, imbalanced data is a key, widespread issue since the goal is usually to find one matching user identity pair out of over a thousand or ten thousand nonmatching user identity pairs, and the number of matching and nonmatching pairs is extremely imbalanced. Previous works have discussed the challenges of classification with highly imbalanced UIL data sets and have recommended the precision and recall as reliable evaluation metrics. [3], [17], [18]. However, little effort has been made to solve the imbalanced UIL problem. In a highly imbalanced environment, for n user identity pairs, none or only one of them is a matching pair. Although we want to identify matching pairs, if the traditional binary classifier predicts that all the inputs belong to the majority class (nonmatching pairs), the prediction accuracy may be 90% or even higher. As an inevitable result, a certain proportion of matching pairs will not be identified. Under/oversampling methods [19], [20] can balance the distribution of majority and minority classes, respectively. In addition, cost-sensitive learning [21], [22] can set the penalty minority class value according to the practical environment, which will not change the distribution of data, e.g., we should increase the cost for "not finding" the matching pair of web-browsing behaviors and also the cost of misclassifying the nonmatching pairs into matching pairs. With the proposed Siamese architecture-based UIL (SAUIL) model to reduce the influences of imbalanced data, in this work, we propose new cost-sensitive loss functions for cost-sensitive learning of features and classifier parameters and examine the performance of our model under different imbalance ratios and under sampling proportions. We summarize our key contributions as follows.

1) We propose a Siamese architecture-based UIL model with BLSTM as a subnetwork. The model compares the similarity of two inputs, i.e., two web-browsing behaviors in our case. The web-browsing pair that has a high similarity value will be classified as belonging to the same real person.
2) Finding one matching pair out of thousands of nonmatching pairs represents an extremely imbalanced classification problem. We propose a cost-sensitive loss function that increases the cost of misclassifying the minority class. Then, we propose the cost-sensitive SAUIL (C-SAUIL) model for imbalanced user identity linkage (UIL) problems through visited webpages.
3) For the evaluation, we collected three very large, real-world data sets with distinct characteristics: 1) the data traffic of mobile Internet from a northern city in China; 2) the data traffic of fixed network from a campus in Beijing, China; and 3) the browsing history of BitTorrent (BT) from a campus network in Beijing, China. The experimental results show that the proposed SAUIL model outperforms other deep NN models, and the cost-sensitive SAUIL model further improves the G-mean and F-measure metrics.

The proposed NN-based learning system can find the highest-level feature representations of user interests are extracted by BLSTM and compared with a distance function in the two twin networks that share the same parameters. We also analyze the effect of the proposed loss functions on the backpropagation algorithm by deriving relations for propagated gradients. The source code of the proposed model can be found at https://github.com/fanduo12138/User_Identity_Linkage, along with the data sets of the BT resource.

The remainder of this paper is organized as follows. In Section II, we summarize related studies. Section III describes the definition and preliminary assumptions of our UIL problem. We introduce the details of our SAUIL model and the cost-sensitive C-SAUIL model in Section IV. Then, the experiments and analysis results are presented in Section V. Finally, we conclude our work in Section VI.

II. Related Work

The increasing popularity and diversification of Internet services, such as social networking, instant messaging, forums, online videos, and online e-commerce sites, have made UIL problems receive increasing attention [3]. Usually, the goal of the UIL problem is to answer a simple question: *do two online identities belong to the same real person?* This question can be seen as a classification problem, which classifies the

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

online identities belong to the same real person to the same class. In handling the classification tasks, the two main challenges are extracting representative features and constructing classifiers with good performance. In this section, we will summarize the previous studies according to these two aspects.

A. Extracting Representative Features

For homogeneous or heterogeneous data sources, representative features are selected, extracted, and compared by different methods and models. Distance functions, such as the Jaro–Winkler distance [23] and the Jaccard index [8], term frequency and probabilistic methods, such as the bag-of-words model [17], the term frequency-inverse document frequency (TF-IDF) [24], N-Gram models [4], and topic models [4], [8], can measure the similarity and difference of texts by comparing the terms or extracting the content topics. Users' spatiotemporal trajectories are matched using frequency or probabilistic methods by calculating the weight for different locations [9]–[11], which is very similar to text-matching methods except for these methods consider time as a necessary attribute. Another method that links user online identity between heterogeneous data sets involves mapping features from multiple platforms to a homogeneous space (usually referred to as the latent space) with an embedding-based method [25]–[31] or a projection matrix [17]. Since every package transmitted over the Internet can be recorded in web logs as flows, which are aggregated from packets by five triples (the source IP, destination IP, source port, destination port, and the protocol), finding the online identity of users through web browsing can directly link online behaviors of a real user. With the probabilistic soft logic method, many features in web logs are leveraged to collectively determine whether sets of web logs belonging to the same user [12], [13]. The URL in web logs records the global address of the documents and other resources on the Worldwide Web. Lexical, host, and content information can be obtained for a URL as feature representations [32]. The above work failed to capture the semantic or sequential patterns and handled unseen features without substantial manual feature engineering; therefore, researchers apply deep learning methods to learn representations of multiple levels of abstraction from URLs [33]. Inspired by previous works, we employ a deep learning method to learn the representative features. We propose a Siamese NN-based architecture with a multilayer perceptron (MLP), a convolutional NN (CNN), or BLSTM as a subnetwork to extract features from web-browsing behaviors.

B. Constructing a Classifier With Imbalanced Learning Methods

Many machine learning algorithms can be applied for solving UIL problems, including supervised, semi-supervised, and unsupervised methods [3], [6], [31]. Although matching online identity pairs is very rare among all pairs, to the best of our knowledge, previous works did not handle the imbalanced UIL problem introduced by extremely imbalanced minority (matching online identity pairs) and majority classes (nonmatching online identity pairs) when constructing the classifier. Two basic strategies for addressing imbalanced learning are preprocessing and cost-sensitive learning [34]. Preprocessing methods, such as oversampling and undersampling, will greatly change the distribution of extremely imbalanced data. The ensemble learning algorithms achieve a good performance on the class imbalance problem because they can increase the accuracy of single classifiers by combining several of them, which must be specifically designed [35]. Recently, cost-sensitive algorithms that assume higher costs for the misclassification of minority class [36], [37] have gained increasing attention since they do not change the original distribution of the data or increase the computational complexity. By learning different weights of the model for different classes [37]–[40] or employing new loss functions [41], [42], cost-sensitive deep NNs are proposed.

Inspired by the above works, we attempt to solve the problem of imbalanced learning by minimizing the overall cost on the training data set with a cost-sensitive classification model that employs a Siamese NN architecture. For the UIL problem, the number of matching and nonmatching pairs is extremely imbalanced, which can be well handled by the Siamese NN architecture. With the two newly proposed loss functions, our model, SAUIL, can also operate under a cost-sensitive setting and can achieve a better performance.

III. Problem Definition and Preliminary Assumptions

A. Problem Definition

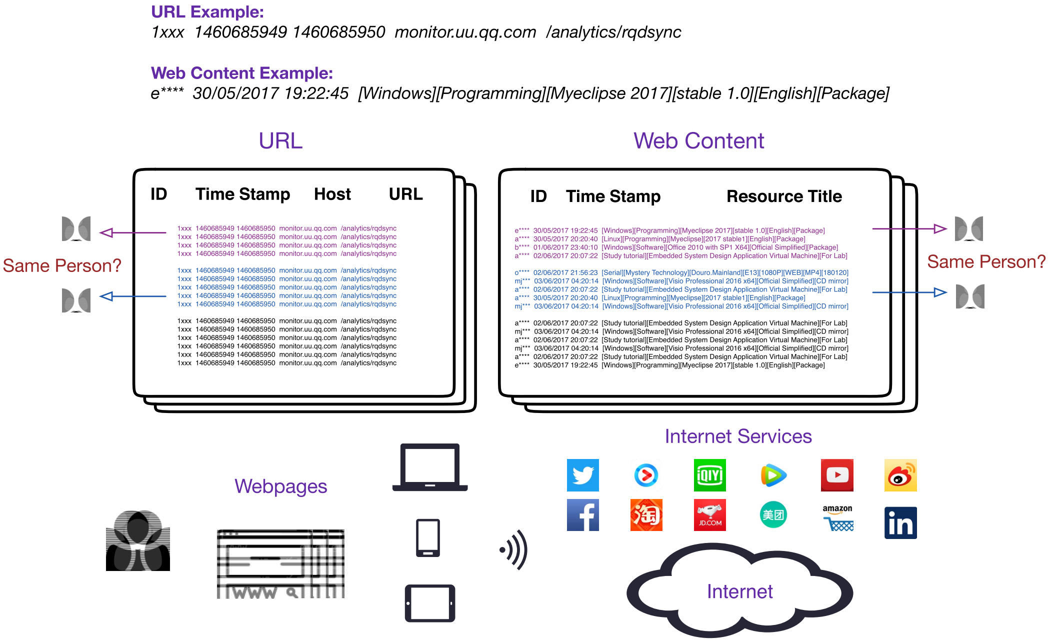
In general, the UIL problem is to link the online behaviors of a real person by extracting the unique properties of users from online activities. In this work, we try to link the web-browsing behaviors of the same user by extracting the intrinsic interest from the web-browsing behaviors, which is recorded by URLs or the content the user browsed, as illustrated in Fig. 1.

Definition 1 (User Web-Browsing Behavior): We define a user's online web-browsing behavior as a triple (u, P, V^k) where $u \in U$ denotes a real user and $U = \{u_1, u_2, \cdots, u_M\}$ is the set of all M users. $P = \{p_1, p_2, \cdots, p_I\}$ is the web-browsing behavior of a user, i.e., a collection of visited webpage sets p_i, where $i = 1, 2, \cdots, I$ for the ith webpage set, which refers to a series of URLs or content instances that a user continuously visited in a dynamic time interval. For each $p \in P$, we obtain a feature mapping $p \rightarrow v$, where $v \in V^k$ is the k-dimensional feature vector representing webpage set p. Then, we have a mapping function $\Phi(u) = V^k$ by learning its inverse function $\Phi^{-1}(V^k) = u$ since we try to infer real users who visited a set of given webpages.

We can observe tens of thousands of webpage sets $p \in OB$, where OB denotes the traces of a user's online browsing behavior, i.e., all webpages the user visited. Our goal is to use a linking function Γ to find all the p that belong to the same user, which is defined below.

Definition 2 (Webpage Set Linking Function): Given all webpage sets in OB, we have a threshold δ. The linking function Γ will calculate the distance of each webpage set pair $(p_i^{t_n}, p_j^{t_{n+1}}) \in OB$ in two adjacent time intervals t_n and

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



URL Example:
1xxx 1460685949 1460685950 monitor.uu.qq.com /analytics/rqdsync

Web Content Example:
*e**** 30/05/2017 19:22:45 [Windows][Programming][Myeclipse 2017][stable 1.0][English][Package]*

```
ID   Time Stamp              Host               URL
1xxx 1460685949 1460685950  monitor.uu.qq.com  /analytics/rqdsync
1xxx 1460685949 1460685950  monitor.uu.qq.com  /analytics/rqdsync
1xxx 1460685949 1460685950  monitor.uu.qq.com  /analytics/rqdsync
1xxx 1460685949 1460685950  monitor.uu.qq.com  /analytics/rqdsync

1xxx 1460685949 1460685950  monitor.uu.qq.com  /analytics/rqdsync
1xxx 1460685949 1460685950  monitor.uu.qq.com  /analytics/rqdsync
1xxx 1460685949 1460685950  monitor.uu.qq.com  /analytics/rqdsync
1xxx 1460685949 1460685950  monitor.uu.qq.com  /analytics/rqdsync
1xxx 1460685949 1460685950  monitor.uu.qq.com  /analytics/rqdsync

1xxx 1460685949 1460685950  monitor.uu.qq.com  /analytics/rqdsync
1xxx 1460685949 1460685950  monitor.uu.qq.com  /analytics/rqdsync
1xxx 1460685949 1460685950  monitor.uu.qq.com  /analytics/rqdsync
1xxx 1460685949 1460685950  monitor.uu.qq.com  /analytics/rqdsync
1xxx 1460685949 1460685950  monitor.uu.qq.com  /analytics/rqdsync
1xxx 1460685949 1460685950  monitor.uu.qq.com  /analytics/rqdsync
```

```
ID     Time Stamp             Resource Title
e****  30/05/2017 19:22:45  [Windows][Programming][Myeclipse 2017][stable 1.0][English][Package]
a****  30/05/2017 20:20:40  [Linux][Programming][Myeclipse][2017 stable1][English][Package]
b****  01/06/2017 23:40:10  [Windows][Software][Office 2010 with SP1 X64][Official Simplified][Package]
a****  02/06/2017 20:07:22  [Study tutorial][Embedded System Design Application Virtual Machine][For Lab]

o****  02/06/2017 21:56:23  [Serial][Mystery Technology][Douro.Mainland][E13][1080P][WEB][MP4][180120]
mj***  03/06/2017 04:20:14  [Windows][Software][Visio Professional 2016 x64][Official Simplified][CD mirror]
a****  02/06/2017 20:07:22  [Study tutorial][Embedded System Design Application Virtual Machine][For Lab]
a****  30/05/2017 20:20:40  [Linux][Programming][Myeclipse][2017 stable1][English][Package]
mj***  03/06/2017 04:20:14  [Windows][Software][Visio Professional 2016 x64][Official Simplified][CD mirror]

a****  02/06/2017 20:07:22  [Study tutorial][Embedded System Design Application Virtual Machine][For Lab]
mj***  03/06/2017 04:20:14  [Windows][Software][Visio Professional 2016 x64][Official Simplified][CD mirror]
a****  02/06/2017 20:07:22  [Study tutorial][Embedded System Design Application Virtual Machine][For Lab]
mj***  03/06/2017 04:20:14  [Windows][Software][Visio Professional 2016 x64][Official Simplified][CD mirror]
a****  02/06/2017 20:07:22  [Study tutorial][Embedded System Design Application Virtual Machine][For Lab]
e****  30/05/2017 19:22:45  [Windows][Programming][Myeclipse 2017][stable 1.0][English][Package]
```

Fig. 1. Illustration of UIL through web browsing.

t_{n+1}, which is considered to be generated by the same real user if the value of the distance between the feature vector of $p_i^{t_n}$, $p_j^{t_{n+1}}$ is larger than δ, that is,

$$\Gamma\left(p_i^{t_n}, p_j^{t_{n+1}}\right) = \begin{cases} 1 & D\left(v_i^k, v_j^k\right) > \delta \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Here, $D : \mathbb{R}^k \rightarrow \mathbb{R}$ is the distance between the k-dimensional feature vector of any two webpage sets $p_i, p_j \in OB$, where D represents the distance function. The condition $\exists q > 0$ such that $D(v_i^k, v_j^k)+q < D(v_i^k, v_{j'}^k)$ must be satisfied. We assume that a user can generate only one webpage set during a time interval since it is very rare for a real person to browse several websites on different equipment and with different IP addresses at the same time, so we do not consider linking the webpage set pairs in the same time interval.

Since the same user shows unchanged intrinsic interests over a long period of time [4], the webpage set linking function should minimize (maximize) the distance between two webpage sets generated by the same (different) actual user or users. Given a set of webpages $P \in OB$ that users visited, to connect the online accessing behaviors of real persons, we get (u, P, V^k) for each user in $U = \{u_1, u_2, \cdots, u_M\}$ by obtaining webpage set linking functions Γ that satisfy

$$\arg\min_{1\le m,n\le M} D(\Phi(u_m), \Phi(u_n)) \tag{2}$$

where u_m and u_n are the same users who visited different p in different time intervals. The webpage set linking functions Γ also satisfy the following:

$$\arg\max_{1\le m,n\le M} D(\Phi(u_m), \Phi(u_n)) \tag{3}$$

where u_m and u_n are different users.

In this paper, the above goal is achieved by learning the hyperparameters in different layers of Siamese NNs by minimizing the expected cost. A loss function will be proposed to weight each possibility that p_i, p_j should be linked or not by estimating the importance of each class.

B. Siamese Neural Networks

Siamese NNs were introduced to solve signature verification as an image-matching problem [43]. Generally, there are two twin networks, and the architecture and weights of each network are exactly the same. Then, the inputs of each network (usually images) are compared by computing the "semantic" distance of the highest-level feature representations [44]. Deep NNs, such as MLP, CNN, and BLSTM, are applied to form twin networks [16], [43], [44]. In this paper, we examine and compare the above three NNs. The overall architecture will be presented in Section IV, followed by detailed experiments in Section V.

IV. Proposed Model

In this section, we will show details about the proposed model. Our approach is to build a trainable framework that maps and compares two different webpage sets so that the distance between two inputs is small if the webpage sets belong to the same real user; otherwise, the distance is large. For the input webpage sets, we adopt a cost-sensitive Siamese NN with a cost-sensitive loss function to compute the similarity of the inputs and to classify the results. Finally, the web-browsing behavior of the same user is identified.

A. Siamese Architecture

Fig. 2 shows the Siamese NN architecture. The webpage sets $p_i^{t_n}$, $p_j^{t_{n+1}}$ are input into each twin network. To learn the

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

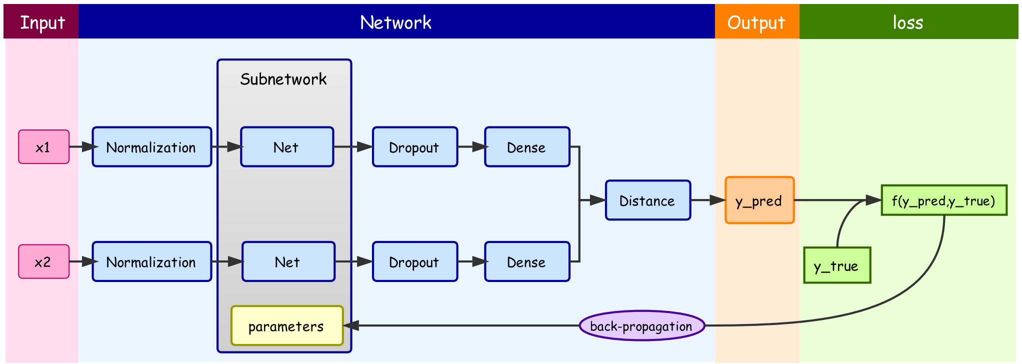


Fig. 2. Siamese NN architecture.

vector representations of the raw URLs or content instances that refer to users' web-browsing behaviors, word vectorization is completed by applying Word2vec [45], [46] with the continuous bag-of-words (CBOW) model and the skip-gram model. Then, the subnetwork that adapts either the MLP, CNN or BLSTM is trained to extract deep feature representations. The twin networks shared all their metrics and parameters, which means that two identical subnetworks are trained. Although the subnetwork is the core of the Siamese NN, it does not have the same constraints as the NN classification algorithm; i.e., it only needs to compute whether two inputs are of the same class or not, instead of knowing which class each input belongs to. In the dropout layer, some units are dropped out randomly with a certain probability p from a Bernoulli distribution during training to reduce overfitting [47]. In the dense layer, each neuron is connected with all the neurons in the previous layer (i.e., fully connected); in this way, the dimensions of the vector can be changed. In our model, the dimensions of vectors in the dense layer are reduced. Then, we apply the Euclidean distance to calculate the distance. To satisfy requirements (2) and (3), we use the exponential function as our distance function D to minimize (maximize) the distance between webpage sets generated from different users (the same user); that is,

$$D(\Phi(\mu_m), \Phi(\mu_n)) = e^{\|\Phi(\mu_m)-\Phi(\mu_n)\|_2-\alpha} \tag{4}$$

where α is a hyperparameter of the proposed model.

B. Subnetwork

In this paper, we consider three deep networks as the subnetwork in the Siamese NN architecture: MLP, CNN, and BLSTM.

1) MLP: An MLP is a class of feedforward artificial NNs with at least three layers of nodes.

2) CNN: CNNs have been widely applied to solve the problem of image representation learning [48]. A CNN consists of one of more convolutional layers for extracting complex features, a fully connected layer on the top, the associated weights and a pooling layer.

3) BLSTM: The LSTM model introduces a new structure, a memory cell, to the traditional recurrent NN (RNN), to reduce exploding or vanishing gradient during the gradient backpropagation phase. BLSTM can efficiently make use of past features (via forward states) and future features (via backward states) for a specific time frame.

C. Cost-Sensitive Learning

Traditional classifiers value the overall prediction precision. In our case, we aim to minimize the overall cost on the training data set with a cost-sensitive classification [49]. More specifically, the minimal expected risk $\mathbb{R}(q_1|\mathbf{x})$ can be expressed as classifying an input vector x into class p

$$\mathbb{R}(q_1|\mathbf{x}) = \sum_{q_2\neq q_1} P(q_2|\mathbf{x})C_{q_1,q_2} \tag{5}$$

where C_{q_1,q_2} is the misclassification cost of misclassifying an instance x that belongs to class q_1 to class q_2. $P(q_2|\mathbf{x})$ is the posterior probability for classifying an input into class q_2. According to (1), for the case of 0–1 classification, we have $C_{q_1,q_2} = 1$, and $C_{q_1,q_1} = 0$. Based on Bayesian decision theory, to reach the minimum overall expectation risk, in our case, we have the ideal classifier that satisfies

$$\underset{q_1}{\arg\min}\,\mathbb{R}(q_1|\mathbf{x}) = \underset{q_1}{\arg\min}\,\mathbb{E}_{(L,Y)}[C(Y,\Gamma(P))] \tag{6}$$

where P is webpage sets, Y is the label vector that only has elements 0 and 1, and $\Gamma(L)$ is the linking function that outputs the label 0 or 1 for the input webpage sets.

To reduce the influence of imbalanced data, we increase the cost of misclassification the minority class by introducing a cost-sensitive loss function. To achieve this goal, we propose a loss function: the cost-sensitive mean distance false error (MDFE). In general, we calculate the loss for the majority and minority classes to increase the cost for the minority class.

D. Cost-Sensitive MDFE Loss Function

The MDFE proposed in this paper is shown in the following:

$$\mathrm{FNE} = \frac{1}{N_n}\sum_{n=1}^{N_n}\left[\left(1-d_n^{(i)}\right)\times\left(y_n^{(i)}\right)^2\right] \tag{7}$$

$$\mathrm{FPE} = \frac{1}{N_p}\sum_{n=1}^{N_p}\left[\left(d_n^{(i)}-0\right)\times\max\left(0,\tau-y_n^{(i)}\right)^2\right] \tag{8}$$

$$l = \mathrm{FNE}+\mathrm{FPE}\times k. \tag{9}$$

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE I

DESCRIPTION OF THE DATA SETS

Type	Source	Duration	The number of users	The number of flows/contents	Geographical area (km^2)
URL	A northern city	68 days	12,874,971	35,571,645,686	473,000
URL	A campus	120 days	2142	550,024,590	0.462
Content	Resource title from BT	2315 days	163,168	18,601,070	-

FNE is the average loss value of the majority class, and FPE represents the average loss value of the minority class. $d_n^{(i)}$ represents the expected value of the ith sample on the nth neuron (tag value), and $y_n^{(i)}$ represents the predicted value of the ith sample on the nth neuron (output value). N_p is the total number of minority classes, N_n is the total number of majority classes, and k is a weight coefficient, which needs to be adjusted according to the data set. τ can be seen as a threshold, and it needs to be adjusted according to the data set, before and after training and testing.

In the network, the relationship between the prediction value $y_n^{(i)}$ and the output of the previous layer is

$$y_n^{(i)} = e^{o_n^{(i)}} \tag{10}$$

and the derivative of $y_n^{(i)}$ with respect to the output of the front layer is

$$\frac{\partial y_n^{(i)}}{\partial o_n^{(i)}} = e^{o_n^{(i)}}. \tag{11}$$

The gradient of the MDFE loss function is

$$\begin{cases} \dfrac{\partial l(d_n^{(i)}, y_n^{(i)})}{\partial o_n^{(i)}} = \begin{cases} -\dfrac{1}{N_n} \cdot 2d_n^{(i)} \cdot \dfrac{\partial y_n^{(i)}}{\partial o_n^{(i)}} & y_n^{(i)} < \tau \\ 0 & y_n^{(i)} \geq \tau \end{cases} & i \in N_n \\ \dfrac{\partial l(d_n^{(i)}, y_n^{(i)})}{\partial o_n^{(i)}} = -\dfrac{1}{N_p} \cdot 2d_n^{(i)} \cdot \dfrac{\partial y_n^{(i)}}{\partial o_n^{(i)}} & i \in N_p. \end{cases} \tag{12}$$

The derivative of the output of the previous layer is

$$\begin{cases} \dfrac{\partial l(d_n^{(i)}, y_n^{(i)})}{\partial o_n^{(i)}} = \begin{cases} -\dfrac{1}{N_n} \cdot 2d_n^{(i)} \cdot e^{o_n^{(i)}} & y_n^{(i)} < \tau \\ 0 & y_n^{(i)} \geq \tau \end{cases} & i \in N_n \\ \dfrac{\partial l(d_n^{(i)}, y_n^{(i)})}{\partial o_n^{(i)}} = -\dfrac{1}{N_p} \cdot 2d_n^{(i)} \cdot e^{o_n^{(i)}} & i \in N_p. \end{cases} \tag{13}$$

The overall loss return function is as follows:

$$\begin{aligned} \frac{\partial l(d_n, y_n)}{\partial o_n^{(i)}} &= \frac{\partial \mathrm{FPE}}{\partial o_n^{(i)}} + \frac{\partial \mathrm{FNE}}{\partial o_n^{(i)}} \\ &= -\frac{k}{N_p} \cdot 2d_n^{(i)} \cdot \frac{\partial y_n^{(i)}}{\partial o_n^{(i)}} \cdot \max(0, \tau - y_n^{(i)}) \\ &\quad -\frac{1}{N_n} \cdot \mathrm{ln}d_n \cdot 2d_n^{(i)} \cdot \frac{\partial y_n^{(i)}}{\partial o_n^{(i)}} \\ &= -\frac{k}{N_p} \cdot 2d_n^{(i)} \cdot e^{o_n^{(i)}} \cdot \max(0, \tau - y_n^{(i)}) \\ &\quad -\frac{1}{N_n} \cdot \mathrm{ln}d_n \cdot 2d_n^{(i)} \cdot e^{o_n^{(i)}}. \end{aligned} \tag{14}$$

Thus, the above loss function can be applied in our model.

V. EXPERIMENTS AND ANALYSIS

A. Data Description

Previous works have mainly linked user identities between different social networking websites. Here, we consider more general application scenarios. We collected data traffic in core networks of Internet service providers (ISPs) of users generated when they connected to the Internet. We capture the Hypertext Transfer Protocol (HTTP) packets to extract the URL that users visited. All packets are aggregated into flows by five triples. In this way, we have the user's anonymous identification, which provides the ground truth, the timestamp of the start and end time of the flow, and the URL the user visited. A URL is a webpage link that users visit, which is only an indirect representation of user web-browsing interests. The most important information is the content the user browsed or is interested in. Therefore, we also collected the resource title of BT from https://bt.byr.cn/ (a BT resource-downloading site on campus networks), the webpages of which listed all the resources (including movie, episode, animation, music, show, game, software, material, sports, documentary) that each user (school student) browsed or downloaded. With these three data sets, the proposed model is examined among users in different regions. In this way, the capability of the model to identify and distinguish users can be fully evaluated. As given in Table I, we collected three data sets in distinct environments in China: the mobile network of a city, the fixed network of a campus, and a BT resource website on a campus network.

On campus, students' interests are relatively similar. Users who live in the city are of different ages, work in different industries, and have different lifestyles. In our experiments, the proposed model is tested by using both data sets. In addition, we further test the model with content that users are interested in by crawling the BT resource title in the browsing history. With the above three real-world data sets, the performance and robustness of the proposed SAUIL and C-SAUIL models are examined under real and diverse environments.

B. Preprocessing

1) Preprocessing of the URL Data: Raw data always have much "noise," such as invalid data and abnormal data, which may introduce bias and should be removed before training the model. Every time we visit a website, many flows will be generated when downloading elements on the webpage from the server. To extract user's browsing interests, dynamic resources, such as Hypertext Markup Language (HTML) [15], are the most important data since they contain the textual content the users browsed. The static resources that provide a function of rendering the webpage rarely change and do not depend on user inputs or preferences, such as URLs ending in *.js, .json, .gif, .png, .bmp, .ico, .rar, .zip, .txt, .flv, .mid, .doc, .ppt, .jpeg, .pdf,.xls, .jpg, .mp3, .wma, .swf, .css*. In our experiment, we remove these resources from the flows. Here, since we only have approximately 2000 valid users (who generated more than 10 000 flows after removing static resources) from the campus-based URL data, we randomly selected 2000 users from the URL data collected from the city and the campus as our original experimental data set.

2) Preprocessing of Web Content: After crawling the contents (resource titles and usernames) from BT resource webpages, we look into the distribution of the number of visited resources by each user. On average, a user visited 27 resources per year. Approximately 90% of users visited fewer than 70 resources in a year.

With the above observation, to extract users' interests from browsed BT resource titles, we selected approximately 20 000 users (top 10%) who visited more than 70 resources to perform the experiments.

3) User Identity Pair Construction: In our experiments, for each user, we divided the URL/content data into two webpage sets by time with the same number of flows/contents. With m users, we have m matching pairs (2 webpage sets in a pair belonging to the same user), $m(m-1)$ nonmatching pairs (2 webpage sets in pair belonging to a different user). Then, for user u_l out of m users, we have a matching webpage set pair $(p_1^{t_n}, p_2^{t_{n+1}})_{u_l u_l}$ that satisfies $\Gamma(p_1^{t_n}, p_2^{t_{n+1}})_{u_l u_l} = 1$. We also have a nonmatching webpage set pair $(p_1^{t_n}, p_2^{t_{n+1}})_{u_l u_k}$ for users u_l and u_k.

C. Evaluation

For evaluation purposes, we choose the optimal hyperparameters and test the performance of the model with different subnetworks. We use the accuracy and F-measure as the classification performance metrics to examine different subnetworks. Finally, we make a detailed comparison of the proposed loss function under different imbalance ratios.

More specifically, we used cross validation to train (eightfold data, 1600 in 2000 matching pairs in the URL data/16 000 in 20 000 matching pairs in content data), verify (onefold data, 200 in 2000 matching pairs in the URL data/2000 in 20 000 matching pairs in the content data), and test (onefold data, 200 in 2000 matching pairs in the URL data/2000 in 20 000 matching pairs in the content data) the model. The proposed model should correctly identify matching and nonmatching pairs. To have matching and nonmatching pairs in specific proportions, for m users with a $p : q$ imbalance ratio, we have m matching pairs and randomly form $(m/p) \times q$ nonmatching pairs.

1) Metrics:

1) *Accuracy*:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{15}$$

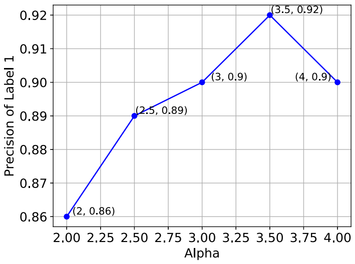


Fig. 3. Grid search for the distance function hyperparameters α.

2) *Recall* (Minority Accuracy):

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{16}$$

3) *Precision*:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{17}$$

4) *G-mean*:

$$\text{G-mean} = \sqrt{\frac{\text{TP}}{\text{TP} + \text{FN}} \times \frac{\text{TN}}{\text{TN} + \text{FP}}}. \tag{18}$$

5) *F-measure*:

$$\text{F-measure} = \frac{(1+\beta^2) \times \text{Recall} \times \text{Precision}}{\beta^2 \times (\text{Recall} + \text{Precision})}. \tag{19}$$

β is a coefficient that adjusts the proportion of precision and recall and is usually set to 1.

2) Choosing the Hyperparameters: Grid search is the most common method to adjust parameters. Here, we find the optimal hyperparameters of the distance function, the dropout layer, and the dense layer by grid search. In Fig. 3, we can clearly see that the model has the highest precision when $\alpha = 3.5$.

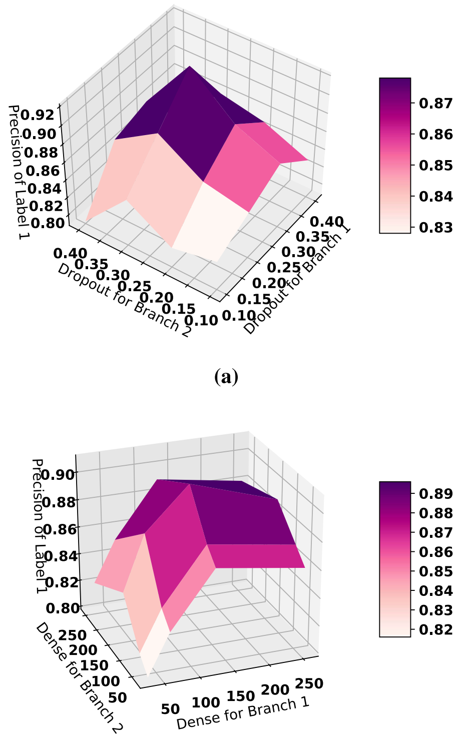
For the dropout layer and the dense layer, the optimal parameters are 0.3 and 128, respectively, as shown in Fig. 4.

3) Subnetwork Comparison: In this section, we test the classification performance of the model with different subnetworks, i.e., MLP, CNN, and BLSTM. The results are given in Table II for the three data sets. For model testing, two URL data sets collected from a city and a campus have 200 users, and the BT resource has 2000 users. The experiments were conducted with a 1 : 3 imbalance ratio; i.e., there are 200 matching pairs and 600 nommatching pairs for the URL data sets and 2000 matching pairs and 6000 nonmatching pairs for the BT resource data sets.

The MLP achieves an average accuracy of 68%, and the CNN predicts that most samples belong to the majority class. BLSTM outperforms the MLP and CNN in terms of the accuracy and the F-measure, especially for the content of the BT resources. There are a very limited number of BT resource types, and students usually maintain a similar preference for BT resources in school, which can be well captured by BLSTM using the browsing history of BT resources for approximately 6 years. Based on the above-mentioned results, we choose BLSTM as the subnetwork of the Siamese NN architecture in the following experiments.

TABLE II
COMPARISON OF THE MLP, CNN, AND BLSTM AS THE SUBNETWORK (CITY/CAMPUS/BT RESOURCE)

Net	Confusion matrix			Accuracy	F-measure
	Matrix	Predicted Negative	Predicted Positive		
MLP	True Negative	453 / 402 / 6000	147 / 298 / 0	0.67 / 0.63 / 0.75	0.39 / 0.40 / 0.00
	True Positive	114 / 101 / 2000	86 / 99 / 0		
	Matrix	Predicted Negative	Predicted Positive		
CNN	True Negative	600 / 592 / 6000	0 / 8 / 0	0.75 / 0.74 / 0.75	- / 0.03 / 0.00
	True Positive	200 / 197 / 2000	0 / 3 / 0		
	Matrix	Predicted Negative	Predicted Positive		
BLSTM	True Negative	567 / 549 / 5959	33 / 51 / 41	**0.79 / 0.80 / 0.99**	**0.42 / 0.43 / 0.98**
	True Positive	138 / 132 / 52	62 / 68 / 1948		



(a)

(b)

Fig. 4. Grid search for the parameters. (a) Probability p in the dropout layer. (b) Number of units in the dense layer.

TABLE III
SUMMARY OF THE HYPERPARAMETERS FOR OUR MODEL

Parameter	Setting
The number of units for BLSTM	200
Probability p in the dropout layer	0.3
The number of units for the dense layer	128

4) Loss Function Comparison: In this section, the performance of the proposed loss function will be examined. Table III presents the values of the hyperparameters for our model. To increase the precision of the minority class (we want to find matching pairs as accurately as possible), the model learns the parameters of BLSTM and the threshold parameter ρ along with the loss function MDFE during training. Here, if the "distance" between the two inputs is smaller than the value of ρ, then the two inputs belong to the same user, and vice versa.

Our model is based on Siamese NNs, which can compare the similarity between two inputs. More specifically, the inputs are mapped into a latent space with a function (BLSTM); then, the similarity can be easily compared by a distance function (the Euclidean distance). During training, our goal is to find the optimal parameters for the model to minimize (maximize) the distance between inputs that belong to the same (different) users. Here, our proposed loss function, cost-sensitive MDFE, will be trained to increase the cost for misclassification of the minority classes. Furthermore, the threshold parameter ρ will be trained along with the loss function. If the distance between two inputs is larger (smaller) than ρ, then the model outputs 1 (0), which means that two inputs are generated by the same user (different users).

With the threshold parameter ρ equal to 0.6 and τ equal to 1.8, we compare the performance of cost-sensitive MDFE with other classic loss functions in Table IV under different imbalance ratios. To save space, we only list the results with a flow data set collected from a city and a campus since the proposed SAUIL model performs very well with a data set of browsed BT resource titles.

In Table IV, the MDFE achieves the best performance in both data sets in terms of the accuracy, precision, and G-mean under different imbalance ratios. Although the value of the precision, recall, F-measure, and G-mean decrease with an increasing imbalanced ratio for almost all the loss functions (we have only one exception: the recall of the hinge loss is 0.99, 1.00, and 1.00 for 1:5, 1:10, and 1:50, respectively), when the imbalance ratios increase, the differences between the MDFE and other loss functions increase; in other words, the MDFE performs better for data with higher imbalance ratios. In terms of the recall, the MDFE performs the best in the campus data set, and we correctly identify all the minority samples in the testing data set collected from the city when using the hinge loss with 1:10 and 1:50 imbalance ratios or

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE IV

COMPARISON OF THE PERFORMANCE FOR DIFFERENT LOSS FUNCTIONS WITH A FLOW DATA SET COLLECTED FROM A CITY/A CAMPUS

	Imb. ratio	Accuracy	Precision	Recall	F-measure	G-mean
Binary cross-entropy	1:5	0.78 / 0.74	0.32 / 0.17	0.33 / 0.14	0.32 / 0.15	0.53 / 0.35
	1:10	0.75 / 0.77	0.12 / 0.02	0.28 / 0.03	0.17 / 0.02	0.47 / 0.15
	1:50	0.74 / 0.77	0.02 / 0.00	0.24 / 0.01	**0.34** / 0.00	0.42 / 0.06
Hinge	1:5	0.17 / 0.13	0.17 / 0.04	0.99 / 0.18	0.28 / 0.06	0.00 / 0.14
	1:10	0.10 / 0.07	0.10 / 0.01	**1.00** / 0.05	0.16 / 0.01	0.00 / 0.06
	1:50	0.02 / 0.03	0.02 / 0.00	**1.00** / 0.04	0.04 / 0.00	0.00 / 0.03
Absolute	1:5	0.69 / 0.77	0.68 / 0.25	**1.00** / 0.19	**0.81** / 0.21	0.28 / 0.40
	1:10	0.33 / 0.80	0.17 / 0.03	0.82 / 0.04	0.28 / 0.04	0.44 / 0.19
	1:50	0.83 / 0.85	— / 0.00	0.00 / 0.00	0.00 / —	0.00 / 0.00
Square	1:5	0.81 / 0.96	0.42 / 0.81	0.39 / 0.46	0.41 / 0.59	0.59 / 0.67
	1:10	0.80 / 0.93	0.15 / 0.67	0.31 / 0.37	0.20 / 0.48	0.51 / 0.60
	1:50	0.79 / 0.90	0.01 / 0.05	0.06 / 0.06	0.01 / 0.05	0.22 / 0.23
MDFE	1:5	**0.93** / **0.93**	**0.94** / **0.83**	0.45 / **0.70**	0.61 / **0.76**	**0.67** / **0.82**
	1:10	**0.94** / **0.94**	**0.91** / **0.70**	0.36 / **0.64**	**0.51** / **0.67**	**0.60** / **0.79**
	1:50	**0.98** / **0.97**	**0.78** / **0.35**	0.15 / **0.61**	0.24 / **0.45**	**0.38** / **0.77**

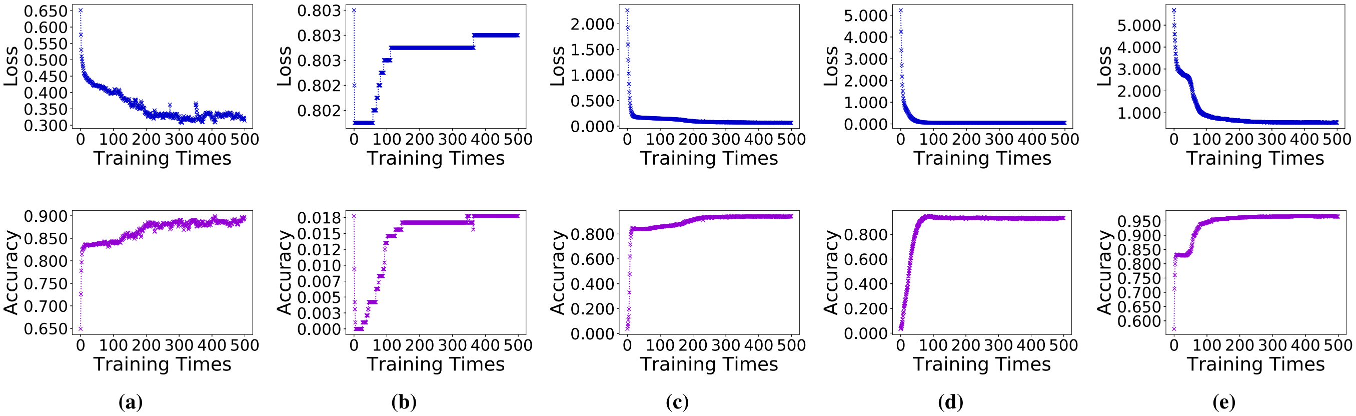


Fig. 5. Comparison of the loss functions for the change in the loss and accuracy values for increasing training time. (a) Binary cross entropy. (b) Hinge. (c) Absolute. (d) Square. (e) MDFE.

the absolute loss with a 1:5 imbalance ratio. The absolute loss, the MDFE, and binary cross entropy have the largest F-measure score under 1:5, 1:10, and 1:50 imbalance ratios for the city data set.

Except for the MDFE, the other loss functions usually suffer from exploding gradients during training. The mathematical reasons behind these trends have been calculated in the Section entitled "Cost-sensitive Learning." We further examine the changes in the loss value and the accuracy for different training times in Fig. 5. During training, the loss and accuracy curves of binary cross entropy are unstable, the loss value of the hinge loss continues to increase, the absolute loss performs very well but obtains very poor classification results in most cases. The square loss has a good performance but achieves a very low precision that is not suitable for our scenario. The loss and accuracy value of the MDFE are relatively stable with increasing training times.

VI. APPLICATION SCENARIO DISCUSSION

In the actual application environment, our model can be trained and work on a monitor deployed on the outlet of the home or company network. The use of our method does not require the user to share data with other individuals or organizations. Collected web logs usually record hundreds of millions of flow records by five triples without a user ID. To distinguish the flow records generated by the same real person, in our previous work [50], [51], we found several features that are widely available in Internet traffic, including the IP address, the "online fingerprint," and the spatiotemporal behavior of the user. The above features are highly discriminative between different users because they usually do not change within a specific time period. Based on these features, we remove part of the noisy data and obtain sets of flow records, each of which is generated by the same real person. Then, the URLs or content visited by the same person can be linked with the proposed C-SAUIL model.

VII. CONCLUSION

In this paper, we propose a Siamese NN architecture-based UIL (SAUIL) model to solve the UIL problem through web browsing. Through a comparison, we employ BLSTM as a subnetwork of the Siamese NN architecture. To further improve the performance of the proposed model by considering the imbalanced UIL problem, we propose a cost-sensitive loss function (MDFE) to increase the costs for misclassifying the minority classes. With a cost-sensitive loss function, we propose the cost-sensitive SAUIL (C-SAUIL) model for an imbalanced UIL problem. Collecting real data sets from different regions and environments, we show that the SAUIL model performs very well in the UIL problem through web browsing and that the C-SAUIL model can improve the overall performance by implementing cost-sensitive classification during training.

The next step will be improving the method and protecting user privacy at the same time. More specifically, when comparing different webpage sets, the URLs/web contents that can reflect an individual's preferences should have higher weights. In addition, the proposed loss function may achieve better minority accuracy (recall) and F-measure by adjusting equations (7–9). How to improve the model by introducing more features without using personal identifiers, as suggested by the GDPR [14], is a key issue as well.

References

[1] CISCO. (2019). *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update*. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/serviceprovider/visual-networking-index-vni/white-paper-c11-738429.html

[2] M. Madden, A. Lenhart, S. Cortesi, and U. Gasser, "Pew Internet and American life project," Pew Res. Center, Washington, DC, USA, 2010.

[3] K. Shu, S. Wang, J. Tang, R. Zafarani, and H. Liu, "User identity linkage across online social networks: A review," *ACM SIGKDD Explor. Newslett.*, vol. 18, no. 2, pp. 5–17, 2017.

[4] Y. Nie, Y. Jia, S. Li, X. Zhu, A. Li, and B. Zhou, "Identifying users across social networks based on dynamic core interests," *Neurocomputing*, vol. 210, pp. 107–115, Oct. 2016.

[5] R. Zafarani and H. Liu, "Connecting users across social media sites: A behavioral-modeling approach," in *Proc. 19th SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2013, pp. 41–49.

[6] Z. Zhong, Y. Cao, M. Guo, and Z. Nie, "Colink: An unsupervised framework for user identity linkage," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 5714–5722.

[7] X. Yu, Y. Sun, E. Bertino, and X. Li, "Modeling user intrinsic characteristic on social media for identity linkage," in *Proc. ACM Conf. Supporting Groupwork*, 2018, pp. 39–50.

[8] S. Liu, S. Wang, F. Zhu, J. Zhang, and R. Krishnan, "Hydra: Large-scale social identity linkage via heterogeneous behavior modeling," in *Proc. SIGMOD Int. Conf. Manage. Data*, 2014, pp. 51–62.

[9] C. Riederer, Y. Kim, A. Chaintreau, N. Korula, and S. Lattanzi, "Linking users across domains with location data: Theory and validation," in *Proc. 25th Int. Conf. World Wide Web*, 2016, pp. 707–719.

[10] H. Wang, Y. Li, G. Wang, and D. Jin, "You are how you move: Linking multiple user identities from massive mobility traces," in *Proc. SIAM Int. Conf. Data Mining*, 2018, pp. 189–197.

[11] D. Kondor, B. Hashemian, Y.-A. de Montjoye, and C. Ratti, "Towards matching user mobility traces in large-scale datasets," *IEEE Trans. Big Data*, to be published.

[12] S. Kim, N. Kini, J. Pujara, E. Koh, and L. Getoor, "Probabilistic visitor stitching on cross-device Web logs," in *Proc. 26th Int. Conf. World Wide Web*, 2017, pp. 1581–1589.

[13] L. Jalali, S. Kim, N. Krishnamoorthy, and R. Biswas, "Using information in access logs for large scale identity linkage," in *Proc. IEEE Int. Conf. Big Data*, Dec. 2018, pp. 2906–2911.

[14] P. Voigt and A. Von dem Bussche, "The eu general data protection regulation (GDPR)," in *A Practical Guide*, 1st ed. Cham, Switzerland: Springer 2017.

[15] *Wikipedia*. Accessed: Apr. 4, 2010. [Online]. Available: https://en.wikipedia.org/wiki/HTML

[16] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *Proc. ICML Deep Learn. Workshop*, 2015.

[17] X. Mu, F. Zhu, F. Lim, E. P. Xiao, J. Wang, and Z.-H. Zhou, "User identity linkage by latent user space modelling," in *Proc. 22nd SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1775–1784.

[18] O. Goga, P. Loiseau, R. Sommer, R. Teixeira, and K. P. Gummadi, "On the reliability of profile matching across large online social networks," in *Proc. 21th SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 1799–1808.

[19] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, 2002.

[20] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker, and G. D. Tourassi, "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance," *Neural Netw.*, vol. 21, nos. 2–3, pp. 427–436, 2008.

[21] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," *Inf. Sci.*, vol. 250, pp. 113–141, Nov. 2013.

[22] V. López *et al.*, "Analysis of preprocessing vs. Cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics," *Expert Syst. Appl.*, vol. 39, no. 7, pp. 6585–6608, 2012.

[23] W. W. Cohen, P. Ravikumar, and S. E. Fienberg, "A comparison of string metrics for matching names and records," in *Proc. Kdd Workshop Data Cleaning object Consolidation*, 2008, pp. 73–78.

[24] Y. Zhang, J. Tang, Z. Yang, J. Pei, and P. S. Yu, "Cosnet: Connecting heterogeneous social networks with local and global consistency," in *Proc. 21th SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 1485–1494.

[25] T. Man, H. Shen, S. Liu, X. Jin, and X. Cheng, "Predict anchor links across social networks via an embedding approach," in *Proc. Int. Joint Conf. Artif. Intell.*, 2016, pp. 1823–1829.

[26] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Me, "Line: Large-scale information network embedding," in *Proc. 24th Int. Conf. World Wide Web (WWW)*, 2015, vol. 2, no. 2, pp. 1067–1077.

[27] L. Liu, W. K. Cheung, X. Li, and L. Liao, "Aligning users across social networks using network embedding," in *Proc. 25th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2016, pp. 1774–1780.

[28] Z. Ma, H. Yu, W. Chen, and J. Guo, "Short utterance based speech language identification in intelligent vehicles with time-scale modifications and deep bottleneck features," *IEEE Trans. Veh. Technol.*, vol. 68, no. 1, pp. 121–128, Jan. 2019.

[29] Z. Ma, Y. Lai, W. B. Kleijn, Y.-Z. Song, L. Wang, and J. Guo, "Variational Bayesian learning for Dirichlet process mixture of inverted Dirichlet distributions in non-Gaussian image feature modeling," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 2, pp. 449–463, Feb. 2019.

[30] W. Zhang, K. Shu, H. Liu, and Y. Wang, "Graph neural networks for user identity linkage," 2019, *arXiv:1903.02174*. [Online]. Available: https://arxiv.org/abs/1903.02174

[31] W. Xie, X. Mu, R. K.-W. Lee, F. Zhu, and E.-P. Lim, "Unsupervised user identity linkage via factoid embedding," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2018, pp. 1338–1343.

[32] D. Sahoo, C. Liu, and S. C. H. Hoi, "Malicious URL detection using machine learning: A survey," 2017, *arXiv:1701.07179*. [Online]. Available: https://arxiv.org/abs/1701.07179

[33] H. Le, Q. Pham, D. Sahoo, and S. C. H. Hoi, "URLNet: Learning a URL representation with deep learning for malicious URL detection," 2018, *arXiv:1802.03162*. [Online]. Available: https://arxiv.org/abs/1802.03162

[34] G. Haixiang *et al.*, "Learning from class-imbalanced data: Review of methods and applications," *Expert Syst. Appl.*, vol. 73, pp. 220–239, May 2017.

[35] M. Jagelid and M. Movin, "A comparison of resampling techniques to handle the class imbalance problem in machine learning: Conversion prediction of Spotify users—A case study," B.S. thesis, KTH, School Comput. Sci. Commun., 2017.

[36] Y. Geng and X. Luo, "Cost-sensitive convolution based neural networks for imbalanced time-series classification," 2018, *arXiv:1801.04396*. [Online]. Available: https://arxiv.org/abs/1801.04396

[37] S. H. Khan, M. Hayat, M. Bennamoun, F. A. Sohel, and R. Togneri, "Cost-sensitive learning of deep feature representations from imbalanced data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 8, pp. 3573–3587, Aug. 2018.

[38] D. Tran, H. Mac, T. Van, H. A. Tran, and N. L. Giang, "A LSTM based framework for handling multiclass imbalance in DGA botnet detection," *Neurocomputing*, vol. 275, pp. 2401–2413, Jan. 2018.

[39] Z. Ma, J. Xie, Y. Lai, J. Taghia, J.-H. Xue, and J. Guo, "Insights into multiple/single lower bound approximation for extended variational inference in non-Gaussian structured data modeling," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published.

[40] Z. Ma *et al.*, "Fine-grained vehicle classification with channel max pooling modified CNNs," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 3224–3233, Apr. 2019.

[41] S. Wang, W. Liu, J. Wu, L. Cao, Q. Meng, and P. J. Kennedy, "Training deep neural networks on imbalanced data sets," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016, pp. 4368–4374.

[42] Y.-A. Chung, H.-T. Lin, and S.-W. Yang, "Cost-aware pre-training for multiclass cost-sensitive deep learning," in *Proc. 25th Int. Joint Conf. Artif. Intell. (IJCAI)*. New York, NY, USA: AAAI Press, 2016, pp. 1411–1417. [Online]. Available: http://dl.acm.org/citation.cfm?id=3060621.3060817

[43] J. Bromley, I. Guyon, Y. LeCun, and E. Säckinger, and R. Shah, "Signature verification using a" Siamese" time delay neural network," in *Proc. Adv. Neural Inf. Process. Syst.*, 1994, pp. 737–744.

[44] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4353–4361.

[45] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. Int. Conf. Learn. Represent.*, 2013, pp. 1–12.

[46] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1188–1196.

[47] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[48] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[49] C. Elkan, "The foundations of cost-sensitive learning," in *Proc. 17th Int. Joint Conf. Artif. Intell.*, 2001, pp. 973–978.

[50] Y. Wu, Q. Lv, Y. Qiao, and J. Yang, "Linking virtual identities across service domains: An online behavior modeling approach," in *Proc. Int. Conf. Intell. Environments (IE)*, Aug. 2017, pp. 122–129.

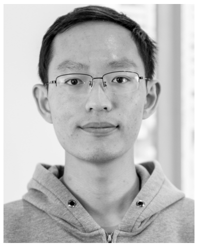
[51] Y. Qiao, Y. Wu, Y. He, L. Hao, W. Lin, and J. Yang, "Linking user online behavior across domains with Internet traffic," *J. UCS*, vol. 24, no. 3, pp. 277–301, 2018.

Yuanyuan Qiao received the B.E. degree from Xidian University, Xi'an, China, in 2009, and the Ph.D. degree from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2014.

She is currently an Associate Professor with the School of Information and Communication Engineering, BUPT. Her current research interests include mobile big data analytics.

Yuewei Wu received the B.E. degree from the Beijing University of Posts and Telecommunications, Beijing, China, where she is currently pursuing the Ph.D. degree with the School of Information and Communication Engineering.

She is engaged in the research of unbalanced data analysis and link prediction.



Fan Duo received the B.E. degree from the Beijing University of Posts and Telecommunications, Beijing, China. He is currently pursuing the master's degree in computer science with the Jacobs School of Engineering, University of California San Diego, San Diego, CA, USA.

He is engaged in the research of unbalanced data analysis and user mobility analysis.

Wenhui Lin received the B.E., M.E., and Ph.D. degrees from the Beijing University of Posts and Telecommunications, Beijing, China, in 2006, 2009, and 2014, respectively.

He is currently a Researcher and an Associate Dean of Technology Research Institute, Aisino Corporation, Beijing, China. His current research interests include traffic measurement and classification, cloud computing, and big data analytics.

Jie Yang received the B.E., M.E., and Ph.D. degrees from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 1993, 1999, and 2007, respectively.

She is currently a Professor and the Deputy Dean of the School of Information and Communication Engineering, BUPT. Her current research interests include broadband network traffic monitoring, user behavior analysis, and big data analysis in Internet and telecommunications.

Dr. Yang was the Vice Program Committee Co-Chair of the IEEE International Conference on Network Infrastructure and Digital Content.