Hindawi
Security and Communication Networks
Volume 2021, Article ID 5533417, 14 pages
https://doi.org/10.1155/2021/5533417

Research Article

User Identification Based on Integrating Multiple User Information across Online Social Networks

Wenjing Zeng,[1] **Rui Tang**,[1] **Haizhou Wang**,[1] **Xingshu Chen**,[1,2] **and Wenxian Wang**[2]

[1]*School of Cyber Science and Engineering, Sichuan University, Chengdu 610065, China*
[2]*Cyber Science Research Institute, Sichuan University, Chengdu 610065, China*

Correspondence should be addressed to Haizhou Wang; whzh.nc@scu.edu.cn

Received 6 January 2021; Revised 24 April 2021; Accepted 11 May 2021; Published 26 May 2021

Academic Editor: Zhe-Li Liu

User identification can help us build more comprehensive user information. It has been attracting much attention from academia. Most of the existing works are profile-based user identification and relationship-based user identification. Due to user privacy settings and social network restrictions on user data crawl, user data may be missing or incomplete in real social networks. User data include profiles, user-generated contents (UGCs), and relationships. The features extracted in previous research may be sparse. In order to reduce the impact of the above problems on user identification, we propose a multiple user information user identification framework (MUIUI). Firstly, we develop multiprocess crawlers to obtain the user data from two popular social networks, Twitter and Facebook. Secondly, we use named entity recognition and entity linking to obtain and integrate locations and organizations from profiles and UGCs. We also extract URLs from profiles and UGCs. We apply the locations jointly with the relationships and develop several algorithms to measure the similarity of the display name, all locations, all organizations, location in profile, all URLs, following organizations, and user ID, respectively. Afterward, we propose a fusion classifier machine learning-based user identification method. The results show that the F1 score of MUIUI reaches 86.46% on the dataset. It proves that MUIUI can reduce the impact of user data that are missing or incomplete.

1. Introduction

With the development of social networks and their diversity, the number of active users on social networks has increased year by year. According to the report of Statista, the number of Facebook active users reached more than 2.7 billion in July 2020, and the number of Twitter active users reached 353 million in July 2020 [1]. People may have accounts on several social networks simultaneously. People can use Twitter to follow the latest developments in their areas of interest, use Facebook to post life trends and keep in touch with friends in life, use LinkedIn to post career information and keep contact with colleagues, and use Foursquare to post locations [2, 3]. If we can match accounts of an individual in different social networks, we can integrate his more comprehensive personal information and draw out their complete friend relationships [4]. This will facilitate social network friend recommendation [5], information diffusion [6], privacy protection [7, 8], community detection [9], etc. [10].

User identification across social networks is also called matching user accounts, user recognition, matching user accounts, user matching, or anchor linking [10]. In recent years, there have been many existing works on user identification across social networks. Most existing works use attributes in the profile to user identification [4, 11–15], such as display name, profile photo, and location. Due to user privacy settings, users may fill in fake information or choose not to fill in. These limitations make these methods quite fragile [16]. Some existing works are relationship-based user identification [17–19]. Relationships have higher discriminability, which is difficult to fake [10]. However, taking into user privacy settings and social network restrictions on data crawl, we may only get part of relationships. This will result in sparse and incomplete relationships. And, a number of existing works also use UGCs to user identification [4, 20].

These methods are usually based on posting time, location, writing style, or similarity of content [4]. However, they may ignore other information contained in the content, such as organizations and URLs. Because of user privacy settings and social network restrictions on data crawl, UGCs may not complete.

For most users, profiles, UGCs, and relationships may be missing or incomplete in real social networks. The features extracted in previous research may be sparse. If more effective features can be extracted from public and available user data, the impact of the above problems can be reduced. Therefore, our paper uses public multiple user information to perform user identification. The main advantages of MUIUI and contributions of our work are

(1) A complete user identification framework: we propose a complete user identification framework MUIUI, which is from data collection to user identification detection. Firstly, we crawl user data from two popular social networks and extract multiple user information from user data, which include profiles, UGCs, and relationships. Then, we extract features from multiple user information. Finally, we employ a fusion classifier to address the user identification problem.

(2) Conducted on popular social networks: this paper focuses on two popular social networks, Twitter and Facebook. We expand the raw dataset, which are those proposed in [21–25], crawled during November 2012 [9]. We screen the users in the raw dataset who are still alive and take them as positive samples. We construct negative samples which display names similar to the display names of half of positive samples. All negative samples and positive samples constitute the dataset used in this paper. We develop multiprocess crawlers to obtain the user data, include profiles displayed in December 2019, and UGCs and relationships published before January 2020, until we reach the limits of the social networks. We can disclose the dataset used in this paper. The MUIUI framework is conducted on this dataset.

(3) Extracted a set of effective features: we use named entity recognition to extract locations and organizations from profile and UGCs and regard them as all locations and all organizations. We use the entity link method to associate the alias of the locations and organizations. We propose methods to calculate the similarity of all locations, the similarity of all organizations, and the similarity of URLs in profile and UGCs. We apply the following relationship jointly with the location in profile to conduct user identification. The experiments prove that the features extracted in this paper are effective for user identification. The experiments also indicate that using multiple user information, we can improve the performance of user identification.

In the rest of this paper, Section 2 presents some related works. Section 3 introduces the basic background and formalizes the problem statement. In Section 4, we describe the user identification framework MUIUI. We do three experiments and compare with three existing works in Section 5. Finally, Section 6 concludes the paper and makes prospects for future work.

2. Related Works

In recent years, there have been much research studies on user identification across social networks. The existing research can be roughly divided into four categories: profile-based user identification, UGC-based user identification, relationship-based user identification, and user identification based on profile and user relationship.

Profile-based user identification only uses profile to identify users. In online social networks, attributes in a profile include the display name, user ID, introduction, location in profile, work education experience, and profile photos. Most research studies use one or more of these attributes. It can prove that these attributes are helpful for user identification. Some existing works only use one attribute for user identification, such as only use display name [11, 13, 26–28], only use profile photos [29], and only use locations [30–33]. These studies prove the feasibility of one attribute to perform user identification. As we know, social networks do not only contain a single attribute. And, applying several attributes jointly can improve the performance of user identification [10]. Li et al. [34] used display names and user IDs to link user identities. Motoyama and Varghese used various attributes, such as display name, location in the profile, age, and email, to link user identities [35]. Due to user privacy settings, users may fill in fake profile information or choose not to fill in. The accuracy of profile-based user identification will decrease.

UGC-based user identification only uses UGCs to identify users. Attributes in a UGC include locations, organizations, time, content, and writing style. Li et al. [4] calculated the similarity of UGCs on spatial, temporal, and content dimensions. Then, they proposed a cascaded three-level machine learning method to solve user identification. Goga et al. [36] used three features extracted from UGCs, such as location attached to UGCs, timestamp, and writing style, to identify users. Because of user privacy settings and social network restrictions on data crawl, UGCs may not complete. The robustness of above identification methods may be poor.

Relationship-based user identification only uses relationships to identify users. Xuan et al. [17] found that users usually maintain a similar circle of friends on different social networks. They use relationships and propose FRUI. Zhang et al. [18] proposed the energy model COSNET by considering the local and global similarities between multiple networks. Zhou et al. [19] sampled the network and learned the vector representation of network nodes. They aligned

anchor nodes through neural networks and link users with dual learning and policy gradient. Some researchers also apply the graph embedding to the user identification. Man et al. [37] used the network embedding method to explore the network structure and identify users through cross-network mapping. Zhou et al. [38] proposed a nonpriority knowledge method FRUI-P based on social relationships. Liu et al. [25] embedded both the following relationship and follower relationship into the network structure to identify users. There are some existing works based on profile and user relationship. Zhang and Yu [39] combined user attributes and network structure to link potential multiple shared entities. Li et al. [10] combined user display name and social network information redundancy to identify users. Zhang et al. [40] extract features from display name, location in the profile, and relationships to identify users. Due to social network restrictions on data crawl, difficulty in obtaining multilevel relationships and the highly dynamic topology of social network [41], relationships will be sparse, incomplete, and unstable.

Relationships can be divided into the following relationships and follower relationships [25]. Due to the openness of social networks, any user can follow other users. A user may not know the person who is following him. Therefore, we only focus on the following relationships. Nowadays, due to user privacy settings and social network restrictions on data crawl, profiles, UGCs, and relationships may be missing or incomplete or fake in real social networks. This paper digged out a set of effective features that is extracted from public and available user data and can reduce the impact of user data which are missing or incomplete.

3. Problem Formulation

Suppose there are two social networks, Twitter and Facebook, represented by G^t and G^f. Use $G^t = \{V^t, E^t\}$ to define social network G^t, where V^t represents the set of all user accounts and E^t represents the set of relationships. User data of user v_i^t include profile, user-generated contents, and relationships. His profile includes display name name_i^t, location loc_i^t, user ID id_i^t, and work education experience we_i^t. His user-generated contents UGC_i^t includes locations UGCL_i^t, organizations UGCO_i^t, and URLs UGCU_i^t. His relationships include the following relationships and follower relationships. The definition of social network G^f is the same as G^t. As shown in Figure 1, we can define user identification across social networks as follows.

User identification: determine whether the user v_i^t in the social network G^t and the user v_k^f in the social network G^f are the same natural person in reality. If they belong to the same natural person, then the user v_i^t and the user v_k^f are called anchor users.

As shown in Figure 2, this paper mainly solves the user identification between two popular social networks, that is, to determine whether two user accounts from two social networks belong to the same natural person. Of course, this method can also be applied to user identification between multiple social networks. The dataset in this paper contains a part of ground truth, that is, anchor link users. We use $A = \{(v_i^t, v_k^f), v_i^t \in V^t, v_k^f \in V^f\}$ to define anchor users. User identification can also be defined as judging whether the user v_i^t and the user v_k^f are anchor users (v_m^t, v_n^f).

4. Model and Solution Framework

The framework proposed in this paper is mainly used for user identification when profiles, UGCs, and relationships are missing or incomplete. Firstly, we introduce the framework as a whole. Then, we specifically introduce the feature extraction methods. Finally, we introduce the fusion classifier machine learning-based user identification method.

4.1. MUIUI Framework. The MUIUI framework includes data crawl and storage module, feature extraction module, and detection module. The MUIUI is shown in Figure 3.

The data crawl and storage module mainly collects user data from Twitter and Facebook and stores it in the MySQL database. This paper uses multiprocess crawlers to crawl user data from Twitter and Facebook. The user data include profile, UGCs, and relationships.

The feature extraction module mainly extracts effective features from multiple user information, which extracts from user data. We obtain fourteen features from a display name and use them as the similarity of display name. The named entity recognition method is used to obtain all locations and organizations from UGCs and profile. We use entity link method to disambiguate and integrate them using the entity link method. We extract all URLs from UGCs and profile. And, extract organizations from the work education experience and combine them with the following relationships to calculate the similarity of the following organizations. We propose several algorithms to measure the similarity of the display name, all locations, all organizations, location in profile, all URLs, following organizations, and user ID, respectively. Combining the above features, a 20-dimensional feature vector is finally obtained.

The 20-dimensional feature vector is input to the detection module to perform user identification. In fact, the detection module uses a fusion classifier. We use the stacking method to fuse three base classifiers which have better performance. The output result of detection module is anchor users or nonanchor users.

4.2. Feature Extraction. Generally, user data contain multiple user information. We can extract several effective features from it. In the following, we exploit multiple user information from network G^t and G^f.

4.2.1. Similarity of Display Name. The display name is closely related to the user. It may not be unique in social networks. At present, some existing works only use the display name as the only attribute for user identification [11, 13, 14]. Compared with other attributes, the display name is easier to obtain. Nevertheless, the user can change the display name at will. The robustness of user identification

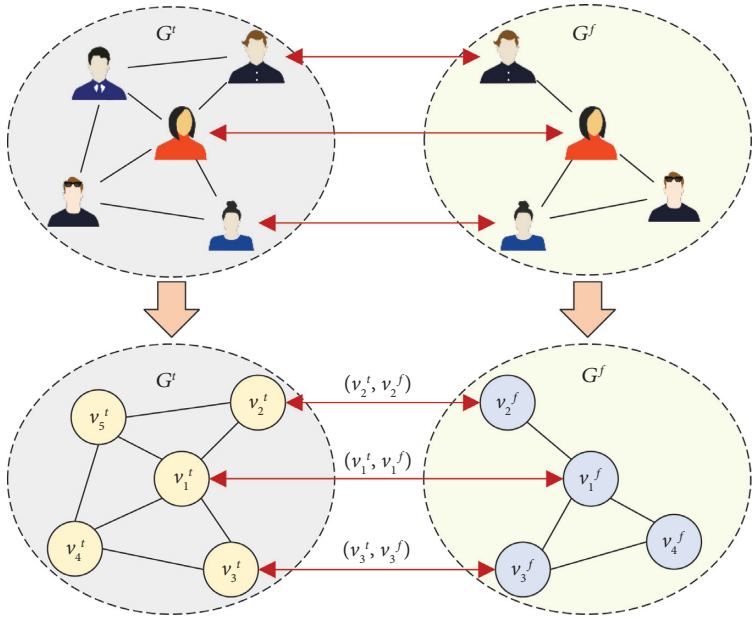


Figure 1: Illustration of user identification across G^t and G^f.

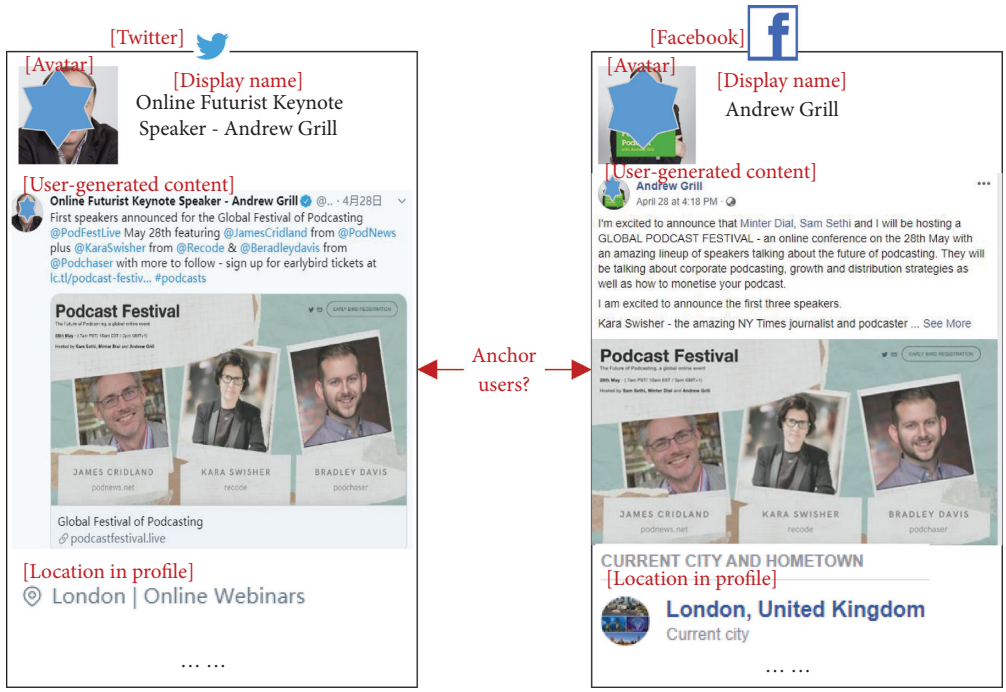


Figure 2: Illustration of user identification across social networks.

based on display name is poor. Li et al. [11] extracted 14 features from the display name. This paper uses their method to obtain features vector X_{ik}^{name} from two display name name_i^t and name_k^f of user v_i^t and v_k^f. We use it as similarity of display name.

4.2.2. Similarity of All Locations and Similarity of All Organizations. In social networks, users may disclose their location in profile, work education experience, and UGCs. The work education experience is filled in by the user and is closely related to the user. The work education experience includes organizations, such as the company where the user works and the school where the user studies. Some social networks include work education experiences directly in the profile (such as LinkedIn and Facebook), and some social networks work education experiences are hidden in the profile (such as Twitter). This paper mainly analyzes the two social networks, Twitter and Facebook. So, for Twitter, we use their introductions as the work education experiences. The content of the UGCs also contains much-hidden

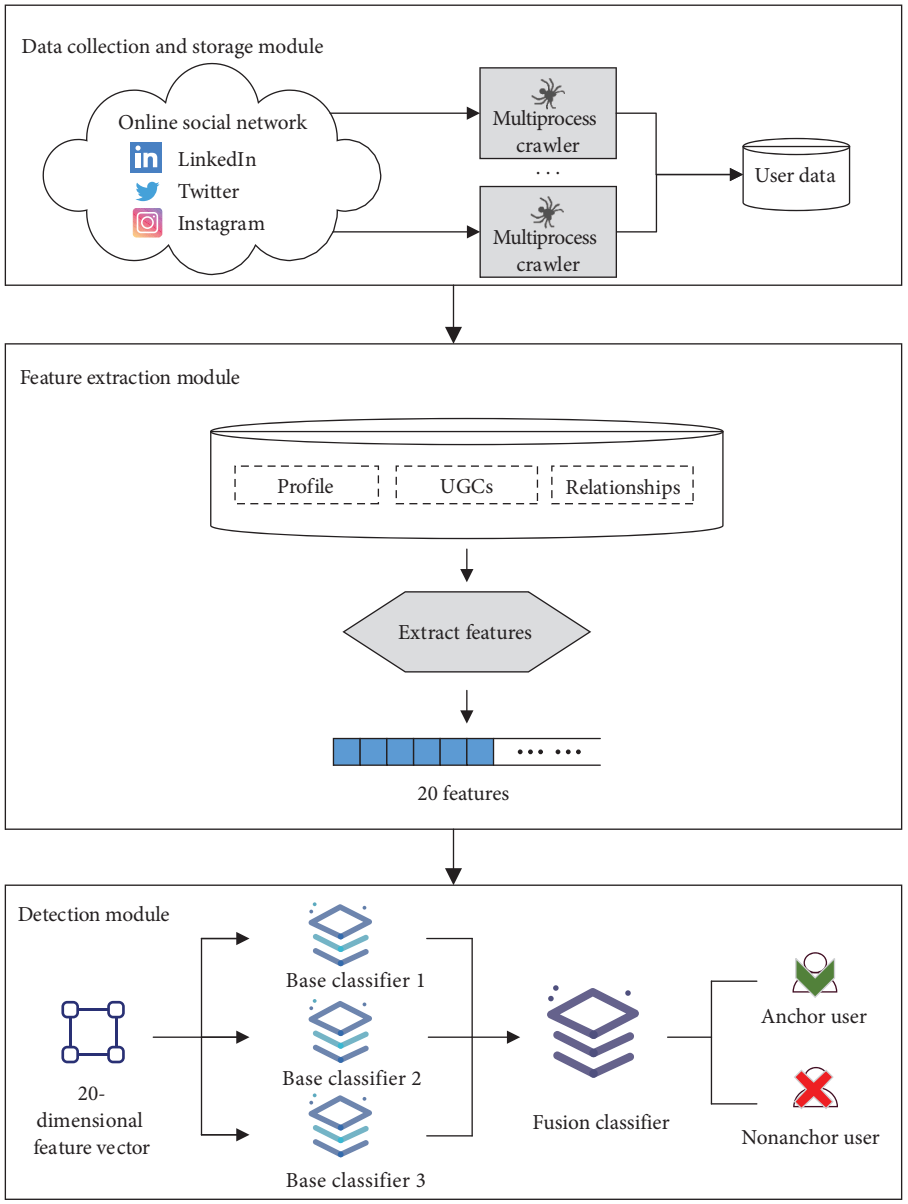


Figure 3: MUIUI framework diagram.

information. For example, locations related to the user, URLs shared by the user, and organizations that the user is concerned about. Named entity recognition can identify named entities from text data. This paper uses named entity recognition to obtain a set of locations and organizations from the content of the UGCs and work education experience. All locations include the location in the profile and the locations involved in the UGCs. Meanwhile, all organizations include the organizations included in the work education experience and the organizations involved in the UGCs.

Since all locations and organizations are closely related to the users themselves, all locations and organizations involved in user public information in different social networks will overlap. Moreover, the more a user mentions the location and organization, the more important it is. The same entity may have many aliases and named entity recognition may also be wrong. The entity link method can solve the above problems. All recognized locations and organizations are mapped to the Wikipedia entry IDs, where names pointing to the same entity are mapped to the same ID. Furthermore, delete entities that do not exist in Wikipedia entries to improve accuracy. This paper uses the named entity recognition method provided by the spacy (https://spacy.io/) library and entity link method provided by the entity link open-source framework Dexter (https://dexter.isti.cnr.it/). The Dexter uses English Wikipedia to implement entity link.

For user v_i^t and user v_k^f, the similarity of all locations and similarity of all organizations can be calculated as follows:

Step 1: for user v_i^t, a set of locations UGCL_i^t and a set of organizations UGCO_i^t are obtained from the content of the UGCs through named entity recognition. Similarly, we obtain a set of locations UGCL_k^f and a set of organizations UGCO_k^f of user v_k^f.

Step 2: for user v_i^t, we obtain a set of locations and a set of organizations from work education experience through named entity recognition. Then, we merge them with the two sets obtained in step 1 to obtain a new set of locations $\mathrm{LOC}_i^t = \{\mathrm{UGCL}_i^t, \mathrm{loc}_i^t\}$ and a new set of organizations $\mathrm{ORG}_i^t = \{\mathrm{UGCO}_i^t, \mathrm{WEO}_i^t\}$. Similarly, we obtain a new set of locations LOC_k^f and a new set of organizations ORG_k^f of user v_k^f.

Step 3: use entity link method to map LOC_i^t and ORG_i^t of user v_i^t into the location ID set LID_i^t and the organization ID set OID_i^t of user v_i^t. Similarly, location ID set LID_k^f and the organization ID set OID_k^f of v_k^f of user v_k^f are obtained.

Step 4: for each $\mathrm{lid}_{im} \in \mathrm{LID}_i^t$ and $\mathrm{lid}_{kn} \in \mathrm{LID}_k^f$, we calculate the weight λ_{im}^t of location ID lid_{im} and the weight λ_{kn}^f of location ID lid_{kn}. For each $\mathrm{oid}_{im} \in \mathrm{OID}_i^t$ and $\mathrm{oid}_{kn} \in \mathrm{OID}_k^f$, we calculate the weight μ_{im}^t of organization ID oid_{im} and the weight μ_{kn}^f of organization ID oid_{kn}.

Step 5: calculating $\mathrm{sim}_{\mathrm{loc}}$ and $\mathrm{sim}_{\mathrm{org}}$ by equations (1) and (2),

$$\mathrm{sim}_{\mathrm{loc}} = \sum_{\mathrm{lid}_{im} \in \mathrm{LID}_i^t, \mathrm{lid}_{kn} \in \mathrm{LID}_k^f} \lambda_{im}^t * \lambda_{kn}^f, \tag{1}$$

$$\mathrm{sim}_{\mathrm{org}} = \sum_{\mathrm{oid}_{im} \in \mathrm{OID}_i^t, \mathrm{oid}_{kn} \in \mathrm{OID}_k^f} \mu_{im}^t * \mu_{kn}^f, \tag{2}$$

where λ_{im}^t is the frequency of lid_{im} in LID_i^t and λ_{kn}^f is the frequency of lid_{kn} in LID_k^f. μ_{im}^t is the frequency of oid_{im} in OID_i^t and μ_{kn}^f is the frequency of oid_{kn} in OID_k^f.

4.2.3. Similarity of Location in the Profile. The location in the profile may be his/her current city or his/her hometown. It is more accurate than the location information extracted from the content of UGCs. Therefore, the similarity of location in the profile is taken as one feature. The profile's location filled in by the same user in different social networks should be closely related [40]. However, there are many aliases for the same location. This paper uses the API provided by pickpoint (https://app.pickpoint.io/) to convert location names into their latitude and longitude. The similarity of location in the profile is calculated based on the latitude and longitude of locations and is expressed by equation (5):

$$ll\left(\mathrm{loc}_k^f, \mathrm{loc}_i^t\right) = \sqrt{\sin^2\left(\frac{\mathrm{lat}_i^t - \mathrm{lat}_k^f}{2}\right) + \cos\left(\mathrm{lat}_i^t\right)\cos\left(\mathrm{lat}_k^f\right)\sin^2\left(\frac{\mathrm{lon}_i^t - \mathrm{lon}_k^f}{2}\right)}, \tag{3}$$

$$d\left(\mathrm{loc}_k^f, \mathrm{loc}_i^t\right) = 2R \times \arcsin\left(ll\left(\mathrm{loc}_k^f, \mathrm{loc}_i^t\right)\right), \tag{4}$$

$$\mathrm{sim}_{\mathrm{home}} = 1 - \frac{d\left(\mathrm{loc}_k^f, \mathrm{loc}_i^t\right)}{C}, \tag{5}$$

where $d(\mathrm{loc}_k^f, \mathrm{loc}_i^t)$ in equation (4) can be measured by equation (3), loc_i^t and loc_k^f represent the location in profile of user v_i^t and user v_k^f, respectively, lat_i^t and lat_k^f are the latitudes of loc_i^t and loc_k^f, respectively, lon_i^t and lon_k^f are the longitudes of loc_i^t and loc_k^f, respectively, and C is a constant, mainly used to normalize the value of $d(\mathrm{loc}_k^f, \mathrm{loc}_i^t)$ (the value of C is 19,860).

4.2.4. Similarity of all URLs. UGCs often include some URLs. These URLs may be the links of UGCs on other social networks, or the links that the user is interested in, or the links related to work education experiences of the user. This paper finds that users may share the same URLs on different social networks. Users may fill in the URL in their profiles, which are often closely related to users. It may be the company web page URL, or the personal web page URL, or homepage URLs of other social networks. Based on these extracted URLs, the similarity of all URLs can be calculated.

We use a method similar to Agarwal's URL extraction methods [12] to extract URLs' set UGCU_i^t and UGCU_k^f from the profile and UGCs, respectively. The calculation method of $\mathrm{sim}_{\mathrm{URL}}$ is shown in equation (6):

$$\mathrm{sim}_{\mathrm{URL}} = \sum_{\mathrm{URL} \in \mathrm{UGCU}_i^t \cap \mathrm{UGCU}_k^f} \gamma^t * \gamma^f, \tag{6}$$

where γ^t and γ^f represent the number of occurrences of the URL in URLs' set UGCU_i^t and URLs' set UGCU_k^f, respectively. URL belongs to the intersection of UGCU_i^t and UGCU_k^f.

4.2.5. Similarity of the following Organizations. Some social networks divide relationships into following relationships

and follower relationships, such as Twitter. Following relationships refer to other users that the target user is following. Meanwhile, follower relationships refer to other users following the target user [25]. Due to the openness of social networks, anyone can become a user's follower. Therefore, we use following relationships and work education experiences to calculate the similarity of the following organizations. The work education experience was introduced in Section 4.2.2. Work education experience includes the organizations where the user works or studies, and these organizations often have their official social accounts in social networks. This paper found that users often follow the official social accounts of organizations that work or study.

This paper mainly analyzes two social networks, Twitter and Facebook. We suppose Twitter is a social network G^t and Facebook is a social network G^f. Because different social networks contain different user information, this paper extracts the organization from the work education experience of Facebook users and obtains the following relationships from Twitter users. Firstly, we extract the homepage URLs from the following users on Twitter and use the entity recognition method to extract the organizations from work education experiences on Facebook. Secondly, we use Google's advanced search method to obtain the official accounts' homepage URLs of the organizations on Twitter (for example, we need to obtain the official account of Apple on Twitter. Google search method is Apple + site: twitter.com). Finally, calculate the similarity of following organizations. For user v_i^t and user v_k^f, the similarity of the following organizations' detailed algorithm is shown in Algorithm 1.

4.2.6. Similarity of User ID. The user ID can uniquely identify a user in the social network. In Twitter and Facebook, the initial value of the user ID is usually automatically generated by the social network, and the initial user ID has a strong correlation with the user's display name. The user can also modify it to a familiar string, but it must be unique. Some research [12] found that user ID can be used for user identification. Therefore, this paper takes the similarity of user ID as one classification feature. The user ID is usually a short string composed of numbers, letters, and underscores so that the string similarity calculation method can be used. This paper uses the Jaro–Winkler algorithm, which is often used to calculate English names' similarity. This algorithm increases the initial characters' weight and makes the string similarity more dependent on the initial part of the string. For user v_i^t and user v_k^f, the calculation method of $\mathrm{sim}_{\mathrm{userid}}$ is

$$d_j = \frac{1}{3}\left(\frac{m}{\left|id_i^t\right|} + \frac{m}{\left|id_k^f\right|} + \frac{m-t}{m}\right), \tag{7}$$

$$\mathrm{sim}_{\mathrm{userid}} = d_j + L \cdot p\left(1 - d_j\right), \tag{8}$$

where id_i^t and id_k^f represent the user ID of user v_i^t and user v_k^f, respectively. m is the number of matching characters and t is the number of transpositions. |id| is the length of user ID and d_j is the Jaro similarity for user ID id_i^t and user ID id_k^f. L is the length of common prefix at the start of the string up to a maximum of four characters and p is a constant scaling factor for how much the score is adjusted upwards for having common prefixes (the value of p is 0.1 in Jaro–Winkler).

4.3. Fusion Classifier. For the same dataset, the effects of different classifiers will also vary. Zhang et al. [40] use logistic regression (LR) and multilayer perceptron (MLP) classifiers to user identification. Liu et al. [42] use support vector machine (SVM) as the model classifier. Zafarani and Liu [43] use logistic regression (LR) as the model classifier. Li et al. [10] use gradient boosting (GB) classifier and tune the parameters of GB to user identification. Li et al. [11] use seven supervised machine learning models and tested them on the training set. Finally, the best model logistic regression with built-in cross-validation (LRCV) is selected as the classifier. These prove that base classifiers can already solve the classification problem well. Li et al. [4] performed ten cross-validations on the classification effect of 10 base classifiers and selected three better base classifiers to construct the fusion classifier. It also proves that the fusion classifier is generally better than the base classifier.

This paper mainly uses a supervised machine learning model to identify anchor users based on the above features. This paper uses 13 classifiers as the base classifiers, including multinomial Naive Bayes (MNB), Gaussian Naive Bayes (GNB), logistic regression (LR), logistic regression with built-in cross-validation (LRCV), support vector machine (SVM), Gaussian process classification (GPC), k-nearest neighbor (KNN), stochastic gradient descent (SGD), multilayer perceptron (MLP), decision tree (DT), random forest (RF), GraBoosting (GraB), and AdaBoost (AdaB). Then, we select three base classifiers with a better performance. Finally, the stacking method is used to fuse three base classifiers to obtain a fusion classifier.

5. Experimental Evaluation

5.1. Experimental Dataset. This paper focuses on two popular social networks Twitter and Facebook. We expanded the raw dataset which are those proposed in [21–25] and crawled during November 2012 [9]. We screened the users in the raw dataset who are still alive and took them as positive samples. We re-crawl 2397 pairs of Twitter and Facebook users in the raw dataset. As a result, 1292 pairs of Twitter and Facebook user accounts were found as still alive. To improve the classifier's performance, 1292 pairs of negative samples are added to the dataset, and half of the negative samples have similar display names to the positive samples. These 2584 pairs of samples are used as the experimental dataset.

We developed multiprocess crawlers to obtain the profiles of the dataset in December 2019 and to obtain UGCs and relationships of the dataset before January 2020, until the limits of the social network. The UGCs can be divided into original and repost. In this paper, we consider the reposted contents to be part of the UGCs,

and the same content reposted multiple times will only be regarded as once. Both Twitter and Facebook users in the dataset are native speakers of English.

5.2. Evaluation Metrics. In the experiments, accuracy, recall, precision, and F1 score are used to evaluate the framework. In this paper, positive samples indicate anchor users, and negative samples indicate nonanchor users.

A confusion matrix is shown in Table 1. TP is the number of samples whose predicted and actual values are both positive. TN is the number of samples whose predicted and actual values are both negative. FN is the number of samples whose predicted is negative but is actually positive. FP is the number of samples whose predicted is positive but is actually negative.

Accuracy (ACC) is the ratio of correct predictions in all samples and is expressed by equation (9):

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}}. \tag{9}$$

Recall (REC) is the ratio of the both predicted and actual are positive samples in all actual samples and is expressed with equation (10):

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{10}$$

Precision (PRE) is the ratio of the both predicted and actual are positive samples in all predicted samples and is expressed by equation (11):

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{11}$$

F1 score is the harmonic mean of precision and recall and is expressed with equation (12):

$$f1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \tag{12}$$

Area under curve (AUC) is area under the ROC curve. AUC can evaluate two-class classifiers. If a classifier has larger AUC, the accuracy of the classifier will be higher.

5.3. Experiments and Analysis. To prove that the MUIUI is an effective user identification framework even when user data are incomplete or missing, this paper makes statistics on the missing and incomplete user data in the dataset, as shown in Table 2. The numerical value in Table 2 is the number of users whose user data are missing or incomplete. Missing information means that the user has not filled in the information or has not disclosed it. Incomplete information means that the user has disclosed and filled in the information, but only part of them can be obtained due to social network restrictions. The false locations are judged by whether location names can be converted into latitude and longitude. If location can be transformed, it is true. Besides, if a user fills in the location is "Earth" or other meaningless nouns, they will also be regarded as false information.

According to the statistics in Table 2, the user data in the dataset used in this paper are missing or incomplete, except for display names. This dataset is crawled from real social networks by multiprocess crawlers. It also proves that user data have varying degrees of missing, falsity, and incompleteness in real social networks. To evaluate the effectiveness of the MUIUI framework, we compare MUIUI with three existing methods: the method proposed by Li [11], the OPL method proposed by Zhang [15], and the ALLEN-LR method proposed by Zhang [40]. The experiments use the dataset introduced in Section 5.1, which has 1292 pairs of anchor users (positive samples) and 1292 pairs of nonanchor users (negative samples). The dataset includes 1881 Twitter users and 1305 Facebook users.

5.3.1. Comparison on Base Classifiers. Use 13 base classifiers to identify users based on dataset introduced in Section 5.1. The base classifiers include multinomial Naive Bayes (MNB), Gaussian Naive Bayes (GNB), logistic regression (LR), logistic regression with built-in cross-validation (LRCV), support vector machine (SVM), Gaussian process classification (GPC), k-nearest neighbor (KNN), stochastic gradient descent (SGD), multilayer perceptron (MLP), decision tree (DT), random forest (RF), GraBoosting (GraB), and AdaBoost (AdaB). These classifiers can be implemented through scikit-learn [44], and all the parameters use their default values. In the experiments, the ratio of positive sample to negative sample is 1 : 1, and the ratio of the training set to the test set is 2 : 1. These 13 base classifiers are tested with the retraining process, and the average results are shown in Figure 4.

According to the results of Figure 4, RF, Grab, and AdaB have the best performance. Grab and AdaB are strong classifiers. A strong classifier is a classifier with higher accuracy, and it works better than weak classifiers. Grab and AdaB belong to strong classifiers and other base classifiers belong to weak classifiers. This is why Grab and AdaB are significantly higher than other classifiers. For RF, if the number of trees (that is, the dimensions of features) is larger, the RF classification performance will be better. The features of this paper reach 20 dimensions, that is, the number of trees is large. So, the RF works better. Therefore, we choose RF, Grab, and AdaB as base classifiers and use the stacking method to construct a fusion classifier as the final classifier.

5.3.2. The Ratio of Positive Sample to Negative Sample. The ratio of positive sample to negative sample in the training dataset may affect user identification framework. In order to choose the ratio of positive sample to negative sample in the MUIUI, the following experiments are based on the ratio of 8 : 1, 6 : 1, 4 : 1, 2 : 1, 1 : 1, 1 : 2, 1 : 4, 1 : 6, and 1 : 8 to train the MUIUI and compare it with the method proposed by Li [11], the OPL method proposed by Zhang [15], and the ALLEN-LR method proposed by Zhang [40]. The results are shown in Figures 5(a)–5(d).

According to the results in Figure 5(a), the accuracy first drops and then rises. Because the number of samples is the smallest at 1 : 1, the accuracy reaches a minimum at 1 : 1. From

Input: the following users FL_i^t of user v_i^t, the work education experiences we_k^f of user v_k^f.
Output: $sim_{org-follow}$.
(1) $UURL_i^t \Leftarrow$ the homepage URLs extracted from the following users FL_i^t
(2) $UURL_k^f = \varnothing$;
(3) $WEORG_k^f \Leftarrow$ the organizations extracted from work education experiences we_k^f by named entity recognition
(4) **for each** $weorg_{kn}^f \in WEORG_k^f$ **do**
(5) $uurl_{kn} \Leftarrow$ the official account's homepage URL of $weorg_{kn}^f$ on twitter obtained by using Google's advanced search method
(6) $UURL_k^f = UURL_k^f + uurl_{kn}$;
(7) **end**
(8) $sim_{org-follow} = |UURL_i^t \cap UURL_k^f|$

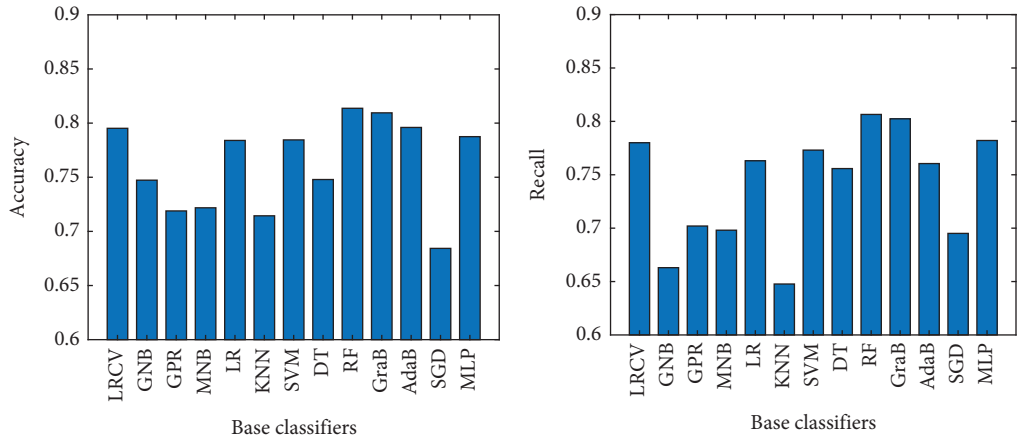
Algorithm 1: Similarity of following organizations.

Table 1: Illustration of confusion matrix.

Actual values	Predicted values	
	Positive samples	Negative samples
Positive samples	TP	FN
Negative samples	FP	TN

Table 2: Statistics on dataset.

Social network	Missing display name	Missing or false location	Missing user-generated content	Missing relationship	Incomplete relationship
Twitter	0	480	62	219	769
Facebook	0	387	0	518	3



(a) (b)

Figure 4: Continued.

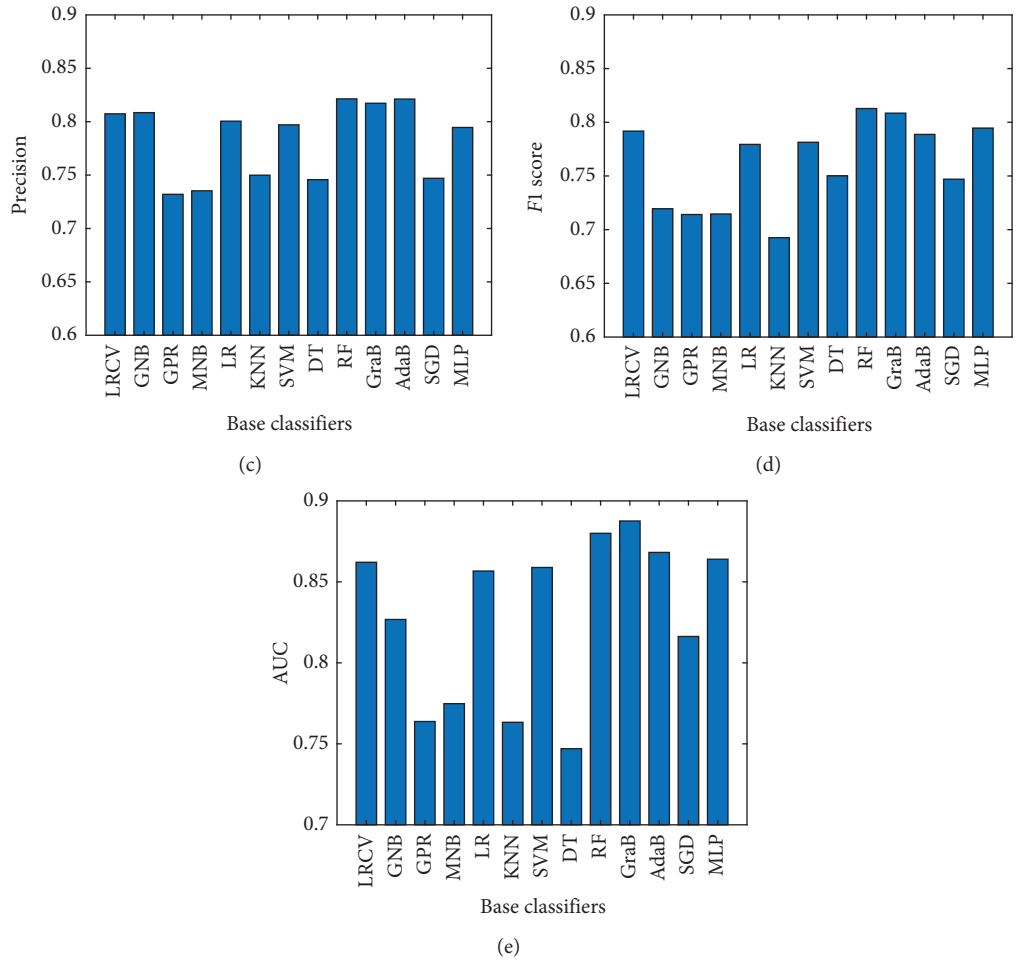


Figure 4: Performance comparison of 13 base classifiers. (a) Accuracy. (b) Recall. (c) Precision. (d) F1 score. (e) AUC.

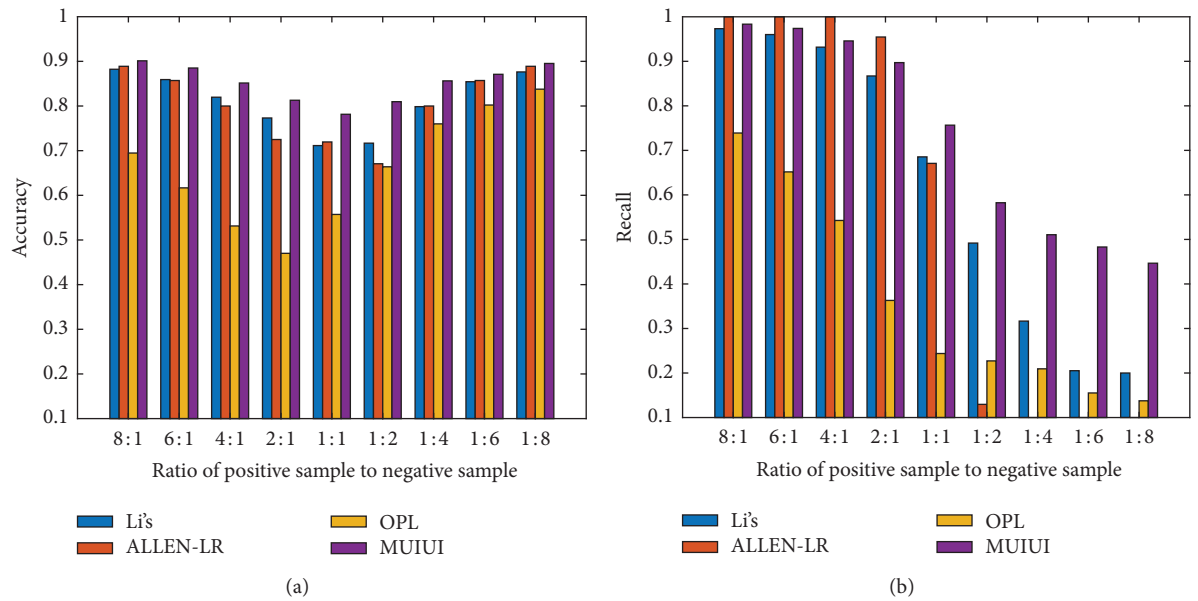


Figure 5: Continued.

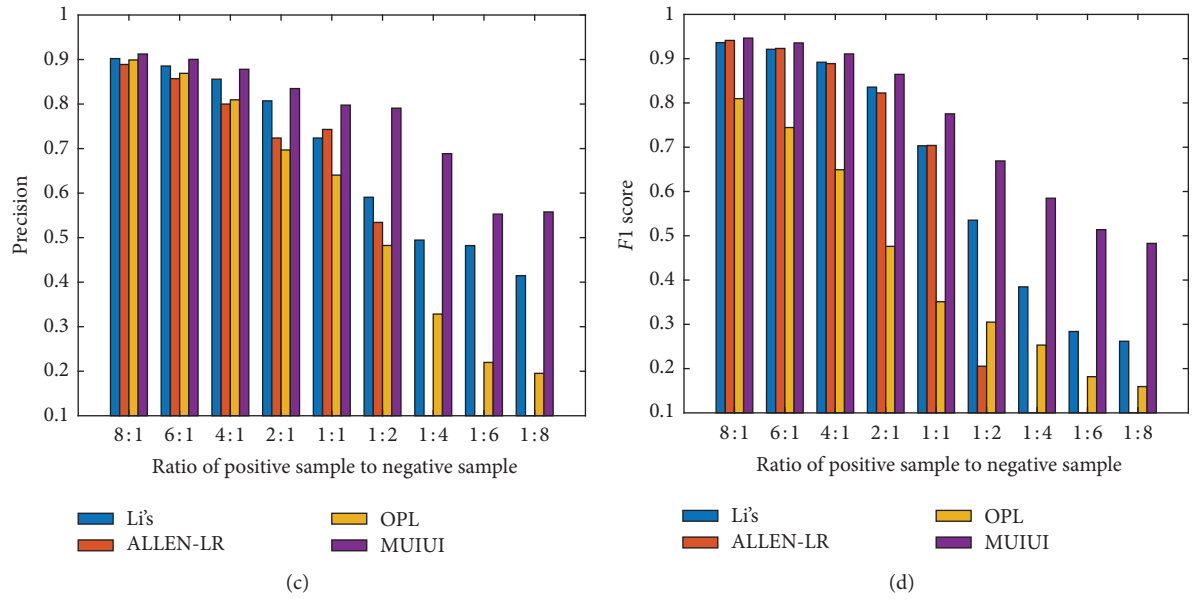


Figure 5: Results with different ratios of positive sample to negative sample. (a) Accuracy. (b) Recall. (c) Precision. (d) $F1$ score.

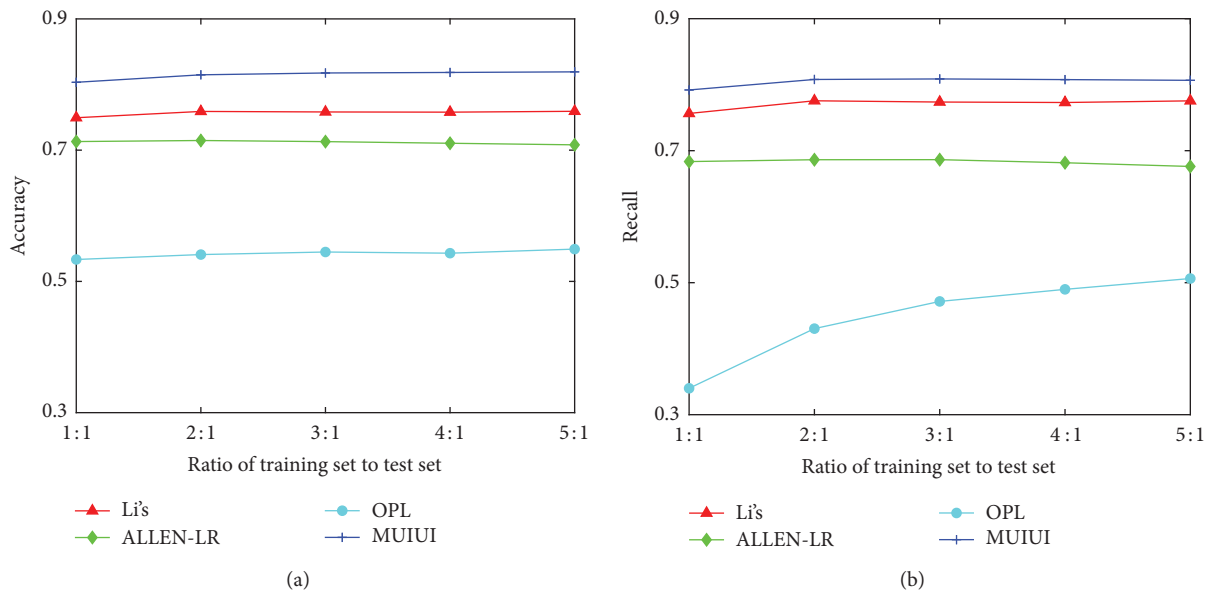


Figure 6: Continued.

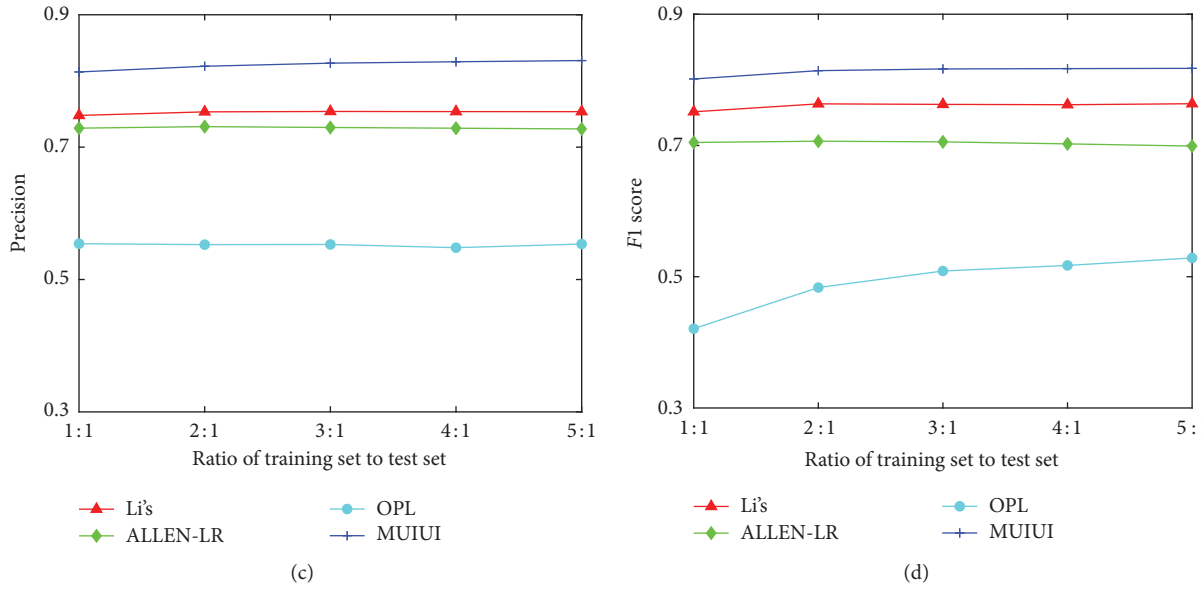


Figure 6: Results with different ratios of training set to the test set. (a) Accuracy. (b) Recall. (c) Precision. (d) *F*1 score.

1 : 1 to both ends, the number of samples increases, and the accuracy is getting higher and higher. The accuracy includes correctly predicted positive and negative samples. The more actual positive samples, the more positive samples are accurately predicted, and it is same for negative samples. The more the samples, the higher the accuracy. Therefore, the accuracy will first decrease and then increase. As shown in Figures 5(b)–5(d), when the proportion of positive samples decreased, the recall, precision, and F1 score also decreased. If training dataset has more positive samples, the classifier will learn more features of the positive samples and predict the positive samples more accurately. Leading to some negative samples are predicted to positive samples.

It can be seen from Figure 5 that the ALLEN-LR method has a higher recall than the method in this paper when positive samples are more than negative samples. However, when negative samples are more than positive samples, the performance of ALLEN-LR drops sharply. When the ratio of positive sample to negative sample is 1 : 4, 1 : 6, and 1 : 8, the recall, precision, and F1 score are almost zero. It shows that ALLEN-LR may judge some negative samples as positive samples. Based on this situation, the F1 score can evaluate the model better. According to Figure 5(d), MUIUI is stable and superior to other methods at different ratios. Because the cost of obtaining positive samples is too high, this paper chooses the ratio of 1 : 1 to construct the dataset.

5.3.3. The Ratio of the Training Set to the Test Set. To more fully illustrate the effectiveness of MUIUI, the following experiments are based on the ratio of the training set to the test set. Different ratio experiments are carried out 100 sampling verifications, and the average of 100 verification results are taken as the final results. According to the results, the accuracy, recall, precision, and *F*1 score of different frameworks are drawn.

Figures 6(a)–6(d) show that the MUIUI has higher indicators than the other three methods under different ratios. At the same time, it can be concluded that the larger the proportion of the training set is, the better the four methods perform.

Li's [11] method only extracts 14 features based on the display name, and there are no missing display names in the dataset. This is the only method without missing user data. The ALLEN-LR method [40] extracts features from the display name, locations in the profile of a user and his/her friends, and the multilayer relationships. It uses the LR classifier to perform user identification. Because the ALLEN-LR method relies heavily on relationships and needs locations in the profile of a user and his/her friends are relatively complete. However, the relationships in our dataset are incomplete, and the location in the profile is partially missing. When the data are partially missing or incomplete, the performance of ALLEN-LR is not ideal. Even if the proportion of the training set increases, it will not help the method. The OPL method [15] proposes methods to complete similarity of the display name, the similarity of profile photo, the similarity of location in profile, the similarity of text in profile, the similarity of URL in the profile, the popularity of the user, and the language user used. These seven features are used for user identification. Because profiles and relationships of some users are missing or incomplete in our dataset, the performance of OPL is also nonideal. It proves that MUIUI can reduce the impact of user data which are missing or incomplete.

6. Conclusion and Future Works

User identification has attracted extensive attention in academic circles, which can be used for friend recommendation, user privacy protection, and advertising recommendation. Due to

user privacy settings and social network restrictions on data crawl, user data may be missing and incomplete in real social networks. The features extracted in previous research may be sparse. In order to solve these problems, we extracted effective features from public and available user data, which can reduce the impact of these problems. Firstly, we developed multi-process crawlers to obtain the latest user data of the dataset. Then, we used named entity recognition and entity linking to obtain and integrate locations and organizations from profiles and UGCs and extracted URLs from UGCs. We developed several algorithms to measure the similarity of the display name, all locations, all organizations, location in profile, all URLs, following organizations, and user ID, respectively. Finally, we proposed a fusion classifier machine learning-based user identification method. We verified the MUIUI framework on the dataset we crawled and the results indicate that the performance is better than that of existing representative works.

Popular social networks LinkedIn and Instagram also contain user data. Our work will be extended to these social networks in the future. We will introduce more effective features into the user identification method, such as user hotspot topics detection, trajectory analysis, and face perception of profile photos. These methods may improve the performance of user identification.

Data Availability

The data supporting this paper are from previously reported studies and datasets, which have been cited. The processed data are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (nos. U19A2081, 61802270 and 61802271). In addition, this work is also partially supported by Joint Research Fund of China Ministry of Education and China Mobile Company (no. CM20200409), and Fundamental Research Funds for the Central Universities (no. 2020SCUNG129).

References

[1] Statista, "Global social networks ranked by number of users 2020," 2020, https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.

[2] R. Tang, S. Jiang, X. Chen, H. Wang, W. Wang, and W. Wang, "Interlayer link prediction in multiplex social networks: an iterative degree penalty algorithm," *Knowledge-Based Systems*, vol. 194, p. 105598, 2020.

[3] J. Zhang, P. S. Yu, and Z.-H. Zhou, "Meta-path based multi-network collective link prediction," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1286–1295, New York, NY, USA, 2014.

[4] Y. Li, Z. Zhang, Y. Peng, H. Yin, and Q. Xu, "Matching user accounts based on user generated content across social networks," *Future Generation Computer Systems*, vol. 83, pp. 104–115, 2018.

[5] S. Huang, J. Zhang, L. Wang, and X.-S. Hua, "Social friend recommendation based on multiple network correlation," *IEEE Transactions on Multimedia*, vol. 18, no. 2, pp. 287–299, 2016.

[6] J. Zhang, P. S. Yu, Y. Lv, and Q. Zhan, "Information diffusion at workplace," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, Association for Computing Machinery*, pp. 1673–1682, New York, NY, USA, 2016.

[7] Y. Qu, S. Yu, L. Gao, W. Zhou, and S. Peng, "A hybrid privacy protection scheme in cyber-physical social networks," *IEEE Transactions on Computational Social Systems*, vol. 5, no. 3, pp. 773–784, 2018.

[8] Y. Qu, S. Yu, W. Zhou, and Y. Tian, "Gan-driven personalized spatial-temporal private data sharing in cyber-physical social systems," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 4, pp. 2576–2586, 2020.

[9] Q. Zhan, J. Zhang, P. Yu, and J. Xie, "Community detection for emerging social networks," *World Wide Web*, vol. 20, no. 6, pp. 1409–1441, 2017.

[10] Y. Li, Z. Su, J. Yang, and C. Gao, "Exploiting similarities of user friendship networks across social networks for user identification," *Information Sciences*, vol. 506, pp. 78–98, 2020.

[11] Y. Li, Y. Peng, W. Ji, Z. Zhang, and Q. Xu, "User identification based on display names across online social networks," *IEEE Access*, vol. 5, pp. 17342–17353, 2017.

[12] A. Agarwal and D. Toshniwal, "Smpft: social media based profile fusion technique for data enrichment," *Computer Networks*, vol. 158, pp. 123–131, 2019.

[13] J. Liu, F. Zhang, X. Song, Y.-I. Song, C.-Y. Lin, and H.-W. Hon, "What's in a name? an unsupervised approach to link users across communities," in *Proceedings of the 6th ACM International Conference on Web Search and Data Mining*, pp. 495–504, Rome, Italy, 2013.

[14] D. Liu, Q. Wu, W. Han, and B. Zhou, "User identification across multiple websites based on username features," *Chinese Journal of Computers*, vol. 38, pp. 2028–2040, 2015.

[15] H. Zhang, M.-Y. Kan, Y. Liu, and S. Ma, "Online social network profile linkage," in *Asia Information Retrieval Symposium*Springer, Berlin, Germany, 2014.

[16] X. Zhou, X. Liang, 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.

[17] Q. Xuan and T. Wu, "Node matching between complex networks," *Physical Review E*, vol. 80, Article ID 026103, 2009.

[18] 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 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1485–1494, Sydney, Australia, 2015.

[19] F. Zhou, L. Liu, K. Zhang, G. Trajcevski, J. Wu, and T. Zhong, "Deeplink: a deep learning approach for user identity linkage," in *Proceedings of the 37th IEEE Conference on Computer Communications*, pp. 1313–1321, Honolulu, HI, USA, 2018.

[20] S. Sajadmanesh, H. R. Rabiee, and A. Khodadadi, "Predicting anchor links between heterogeneous social networks," in *Proceedings of the IEEE/ACM International Conference on*

Advances in Social Networks Analysis and Mining (ASONAM), pp. 158–163, Assam India, 2016.

[21] X. Kong, J. Zhang, and P. S. Yu, "Inferring anchor links across multiple heterogeneous social networks," in *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management*, pp. 179–188, San Francisco, CA, USA, 2013.

[22] J. Zhang, X. Kong, and P. S. Yu, "Transferring heterogeneous links across location-based social networks," in *Proceedings of the 7th ACM International Conference on Web Search and Data Mining, Association for Computing Machinery*, pp. 303–312, New York, NY, USA, 2014.

[23] J. Zhang and S. Y. Philip, "Integrated anchor and social link predictions across social networks," in *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, pp. 2215–2132, Buenos Aires, Argentina, 2015.

[24] J. Zhang, "Social network fusion and mining: a survey," 2018, https://arxiv.org/abs/1804.09874.

[25] L. Liu, W. K. Cheung, X. Li, and L. Liao, "Aligning users across social networks using network embedding," in *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, pp. 1774–1780, Palo Alto, CA,USA, 2016.

[26] R. Zafarani and H. Liu, "Connecting corresponding identities across communities," in *Proceedings of the Third International AAAI Conference on Weblogs and Social Media*, San Jose, CA, USA, 2009.

[27] D. Perito, C. Castelluccia, M. A. Kaafar, and P. Manils, "How unique and traceable are usernames?" in *Proceedings of the 11st International Symposium on Privacy Enhancing Technologies Symposium*, pp. 1–17, Waterloo, ON, Canada, 2011.

[28] Y. Li, Y. Peng, Z. Zhang, M. Wu, Q. Xu, and H. Yin, "A deep dive into user display names across social networks," *Information Sciences*, vol. 447, pp. 186–204, 2018.

[29] A. Acquisti, R. Gross, and F. D. Stutzman, "Face recognition and privacy in the age of augmented reality," *Journal of Privacy and Confidentiality*, vol. 6, pp. 1–20, 2014.

[30] C. Riederer, Y. Kim, A. Chaintreau, N. Korula, and S. Lattanzi, "Linking users across domains with location data: theory and validation," in *Proceedings of the 25th International Conference on World Wide Web*, pp. 707–719, Montreal, Canada, 2016.

[31] W. Chen, H. Yin, W. Wang, L. Zhao, W. Hua, and X. Zhou, "Exploiting spatio-temporal user behaviors for user linkage," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, Association for Computing Machinery*, pp. 517–526, New York, NY, USA, 2017.

[32] X. Gao, W. Ji, Y. Li, Y. Deng, and W. Dong, "User identification with spatio-temporal awareness across social networks," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, Association for Computing Machinery*, pp. 1831–1834, New York, NY, USA, 2018.

[33] W. Chen, H. Yin, W. Wang, L. Zhao, and X. Zhou, "Effective and efficient user account linkage across location based social networks," in *Proceedings of the 34th IEEE International Conference on Data Engineering*, pp. 1085–1096, Paris, France, 2018.

[34] Y. Li, Y. Peng, Z. Zhang, H. Yin, and Q. Xu, "Matching user accounts across social networks based on username and display name," *World Wide Web*, vol. 22, no. 3, pp. 1075–1097, 2019.

[35] M. Motoyama and G. Varghese, "I seek you: searching and matching individuals in social networks," in *11th ACM International Workshop on Web Information and Data Management (WIDM 2009)*, pp. 67–75, Hong Kong, China, 2008.

[36] 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 22nd International Conference on World Wide Web*, pp. 447–458.

[37] T. Man, H. Shen, S. Liu, X. Jin, and X. Cheng, "Predict anchor links across social networks via an embedding approach," in *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, vol. 16, pp. 1823–1829, Palo Alto, CA, USA, 2016.

[38] X. Zhou, X. Liang, X. Du, and J. Zhao, "Structure based user identification across social networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 6, pp. 1178–1191, 2018.

[39] J. Zhang and P. S. Yu, "Pct: Partial co-alignment of social networks," in *Proceedings of the 25th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee*, pp. 749–759, Montreal, Canada, 2016.

[40] Y. Zhang, J. Fu, C. Yang, and C. Xiao, "A local expansion propagation algorithm for social link identification," *Knowledge and Information Systems*, vol. 60, no. 1, pp. 545–568, 2019.

[41] S. Peng, G. Wang, Y. Zhou et al., "An immunization framework for social networks through big data based influence modeling," *IEEE Transactions on Dependable and Secure Computing*, vol. 16, no. 6, pp. 984–995, 2019.

[42] 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 2014 ACM SIGMOD International Conference on Management of Data*, pp. 51–62, Snowbird, UT, USA, 2014.

[43] R. Zafarani and H. Liu, "Connecting users across social media sites: a behavioral-modeling approach," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 41–49, Chicago, IL, USA, 2013.

[44] F. Pedregosa, G. Varoquaux, A. Gramfort et al., "Scikit-learn: machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.