Across Online Social Network User Identification Based on Usernames

Zijian Li, Di Lin(✉), and Peidong Li

University of Electronic Science and Technology of China, Chengdu, Sichuan, China
202021090329@std.uestc.edu.cn, lindi@uestc.edu.cn

Abstract. Cross social network user identification aims to identify the same entity on various online social networks to enhance the completeness and accuracy of the persona. There are three broad categories of cross-social network user identification methods: user identification on account of basic user information, user identification on the basis of network topology graphs, and user identification based on the user's origin. This paper analyzes users' display names from different social networks to determine whether they are the same person. The process consists of three steps: first, we obtain information about users and bring their display names from social networking sites. Secondly, we analyze the user's name, get a series of values from the user's name through similarity calculation methods, and match the similarity. We perform similarity matching on the real dataset by using some classification models. Our model performs well, with F1 values reaching 97.07%, 94.65%, and 92.05% for the three datasets, respectively. This paper can provide a high-quality dataset for downstream NLP tasks of high research significance and value.

Keywords: Across Social Network · Similarity · Feature Extraction

1 Introduction

With the spread of computers and the rapid Internet development, social media sites have become increasingly popular and diverse. More people are participating in various social networks to enjoy the convenient and exciting services they offer. In the real world, user information among social networks is isolated from each other due to user privacy leakage. It is difficult for online social network operators to access other users' information on other social networks, which readily forms data silos and makes it difficult to make a complete character portrait of users. Therefore, it is necessary to recognize users' accounts on different social networks, maximize the integration and perfection of user information, mine the user's social data, and build a relatively complete portrait image base for each user, Provide knowledge assurance for downstream tasks such as recommendation systems, entity alignment, etc.

Account association encounters many challenges: there are few public datasets, and there is still no unified public and comprehensive association dataset that allows all algorithms to make uniform comparisons; privacy breaches are involved, and crawling

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2023
Published by Springer Nature Switzerland AG 2023. All Rights Reserved
X. Jiang (Ed.): MLICOM 2022, LNICST 481, pp. 117–127, 2023.
https://doi.org/10.1007/978-3-031-30237-4_11

data from websites will encounter rate and permission limitations, making it difficult to crawl complete information about users' friendships and user attributes; the number of user association pairs required for the training set may be small, and the distribution may be very uneven, all of which affect the accuracy of the association.

In this paper, we use our user information collected from different social networks as a dataset to classify virtual users from various websites using similarity calculation methods and similarity matching methods to find the same person who points to the natural person in different social networks. Meanwhile, our job only analyzes the display name of users, making obtaining data less challenging, and the experiment shows that the model performs well on the actual dataset.

The remainder of this article is as follows: Section 2 describes some related work, Sect. 3 introduces the acquisition of the dataset and some methods for similarity calculation, and Sect. 4 presents the model for user identification. Finally, the conclusion is given in Sect. 5.

2 Related Work

2.1 Social Network

The term social networking was first used by J. A. Barnes in 1954 [1]. Social networking originated from online social networking [2], which started with E-mail. The Internet is intrinsically a network among computers; the early E-mail figures out the problem of remote mail transmission, and it is still the most famous application. It is also the starting point for online social networking. Behavior-Based Safety (BBS) is a step further. The "group" and "forwarding" normalized, theoretically achieving the function of posting information and discussing topics with all people. Behavior-Based Safety (BBS) to the network social step forward from the simple p2p communication cost lowered to the price of p2p communication. Instant messaging and blogs are likely improved versions of the two aforementioned social tools. The former improves the immediate effect (speed of transmission) and the ability to communicate simultaneously (data processing). The latter began to reflect psychological and sociological theories. That is, the distribution of information nodes start to reflect a stronger individual sense, as this can aggregate distributed information in the temporal dimension to become the "image" and "personality" of the distribution of information node. Commonly used social networks include Weibo, Foursquare, Twitter, Facebook, etc.

2.2 User Identification Based on Online Social Networks [3]

Current techniques for user identification across online social networks have two steps.

Similarity Calculation of User Account. User account similarity calculation methods include similarity calculation based on user attribute information [4], similarity calculation based on network topology map information, and similarity calculation based on user behavior. For the analysis of user attributes, user attributes contain efficient identification information such as user hobbies, occupation, age, etc. The identification methods based on user relationships and user-generated content are difficult to quantify

and model, compared with user attributes, which are relatively easier to "get started," so this method is the most widely studied and popular among scholars. Since user data can be stored as strings [5], we can obtain the similarity values of the corresponding user data items by computing the similarity between the string sequences. The related similarity calculation methods include Levenshtein Distance, Dice Coefficient, Jaro Distance, Named Match Distance [6], etc. For network topology similarity calculation mainly depends on the degree of similarity between different network topologies between two nodes to decide whether these accounts point to the same entity. In other words, the more similar the nodes of the network topology are, the higher possibility that the person of two accounts refer to the same in real life. Followership and followings can be easily gained by open application programming interfaces (APIs) in social networks. Similarity matching between nodes contains methods such as Common Neighbor, Adamic-Adar indicator, resource allocation algorithm, etc. User behavior information, also known as user-generated content (UGC), is shared, exchanged, and posted by users using social networks. The UGC will play an essential role in user identification if it can be fully utilized. The similarity of behavioral information between different social accounts is then used to determine whether the user's identity matches. Related methods include Latent Semantic Analysis and Latent Dirichlet Allocation.

Matching User Account [7]. After the similarity values of two accounts from different social networks are obtained using the above methods, matching among accounts can be implemented using relevant matching algorithms. Mainly including the Kuhn-Munkres algorithm [8], stable marriage matching [9], and ranking-based cross-matching (RCM) [10]. RCM aims to find more matching pairs accurately [11]. Thus the identification of seed set users is decomposed into a distributed iterative process. The seed is a collection of one or more pairs of accounts that are known in advance to be pointing to the same entity in reality. Each iterative process has three stages: the selection of accounts, account matching, and cross matching. Matching funds are checked in each iteration, and the iteration ends when there is no user-matched pair is identified. However, if the precision or accuracy of the seed is too low, it can lead to a reduction in the effectiveness of the model. After the accounts are matched, the feasibility and robustness of the algorithm are determined by measuring the evaluation metrics such as accuracy rate [12] and recall rate [13].

3 Data Collection and Feature Extraction

3.1 Dataset

There is no complete public dataset so far, so we have to use crawlers to get the relevant dataset. Usually, most OSN sites do not allow a specific IP to generate too many requests simultaneously, so we used a distributed crawler system to obtain the data. The whole crawling structure is rough as follows (Fig. 1):

The user information of Foursquare may contain one or more links on Facebook, Twitter, and Instagram of the user. We get these links through Foursquare and then go to the corresponding websites to get the user information through these links. Since

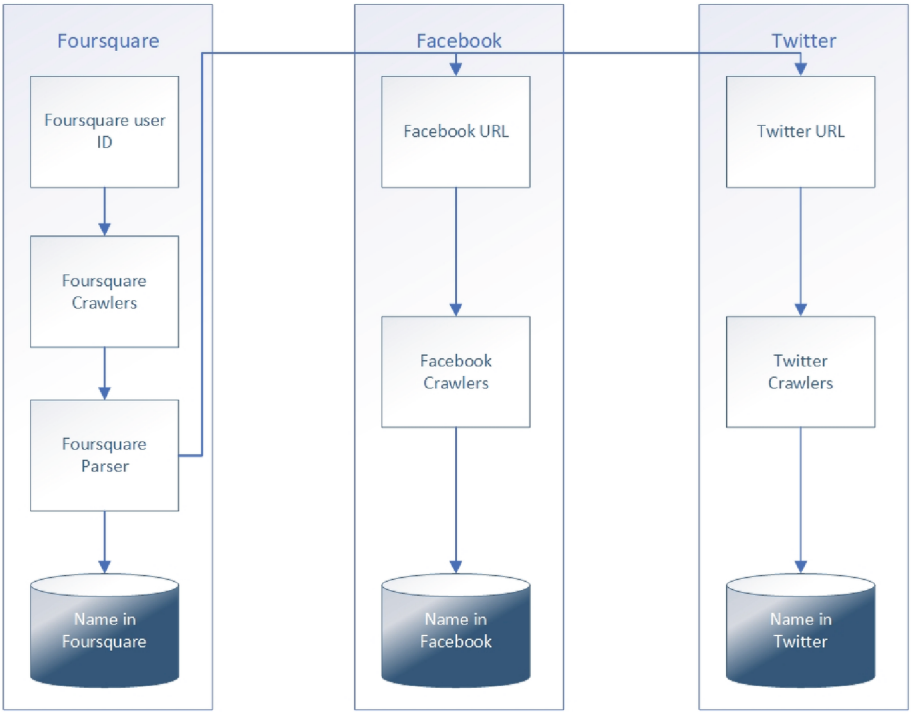

Fig. 1. Procedure of data collection

there were too few users with Instagram links in Foursquare, we only obtained users' Facebook and Twitter links. We got 408,723 user information, among which 225,173 users included Foursquare and Facebook account information, 130,584 users included both Foursquare and Twitter account information, and 9,996 users had both Twitter and Facebook account information. The relevant information is shown in Table 1, and each dataset contains two columns, which are the names of the users of the two platforms.

Table 1. Number of instances in datasets

Name of dataset	Size
FS-TW	225,173
FS-FB	130,584
FB-TW	9,996

The data in Table 1 we call positive samples. At the same time, we construct the negative sample dataset relative to it according to the size of the dataset, and the negative sample dataset is formed by putting different people from different platforms together. It should be noted that we assume that the user's name is unique in the platform and do

not include the case of renaming. At the same time, the language of user names may vary widely, considering the possible inaccuracy when the characters are translated. We only obtain information about users whose user names are in English.

3.2 Feature Extraction

As we all know, user names contain rich meanings. We can calculate the similarity of names by some similarity calculation methods to provide rich information for the subsequent similarity matching. The similarity calculations we use include the following methods (Table 2):

Table 2. Similarity calculation methods

No.	Name of method
1	Average of Best Match
2	Max of Best Match
3	Longest Common Substring Similarity
4	Longest Common Substring Comparing Minimum Length
5	Longest Common Substring Comparing Maximum Length
6	Longest Common Subsequence Comparing Minimum Length
7	Longest Common Subsequence Comparing Maximum Length
8	Edit Distance
9	Normalized Edit Distance
10	Longest Common Substring
11	Longest Common Subsequence
12	Edit Distance Comparing Minimum Length
13	Edit Distance Comparing Maximum Length
14	Normalized Edit Distance Comparing Minimum Length
15	Normalized Edit Distance Comparing Maximum Length
16	Jensen-Shannon Divergence of Alphabet Distribution

The relevant methods are calculated as follows:

Max (Average) of Best Match. Usually, the user's name includes < [first name], [middle name], [last name] >. If the user's name is only taken as a whole, then much information will be lost, so we calculate the similarity of the users' display names by the Average (Max) of Best Match. The specific method is as follows: In the first step, the user's display name is split into words, and for each word, their similarity is calculated first, and the formula for calculating the similarity is as (1).

$$Sim(s_1,\ s_2) = \frac{len(lcs(s_1, s_1))}{(len(s_1) + len(s_2))} \tag{1}$$

where $len(s)$ refers to the length of s and $lcs(s_1, s_2)$ refers to the length of the longest common substring of s_1 and s_2. Each user name can be split into one or more words so that two arrays Arr_1 and Arr_2 can be obtained.

For $w_i \in Arr_1$ and $w_j \in Arr_2$, the second step uses (1) to calculate their similarity to obtain matrix $A = \{a_{ij}\}$, where $a_{ij} = Sim(s_1, s_2)$.

The maximum value is calculated and recorded for matrix A in the third step. Then, delete the row and the column where the maximum value is located, and the third step is repeated until the size of matrix A becomes zero. And then, we get a list containing the maximum value of matrix A after each change, and the maximum value is the result of the Max of Best Match, and the value of the Average of Best Match is the result of the change of the matrix A. The value of the Average of Best Match is obtained by averaging it.

Longest Common Substring Similarity. Longest Common Substring Similarity aims to compute the longest substring of two strings [14]. The calculation is as follows: In the first step, a counter called cn is defined to indicate the number of comparisons and initialized to zero. In the second step, for $name_1$ and $name_2$, unless the length of the longest common substring of the two strings is zero, the following calculation is performed to calculate the length of the longest common substring of both and record it in a list, and the counter is added by one. And then, for $name_1$ and $name_2$, we delete the substring and repeat the second step until the length of the longest common substring of both is zero. Step two is repeated until the length of the longest common substring is zero. In the third step, we determine whether cn is zero. If it is zero, set it to 1, which means there is no common substring between two strings, and then sum the list to get the total and calculate the Longest Common Substring Similarity of $name_1$ and $name_2$, the calculation formula is as follows (2).

$$Sim_{LCS}(name_1, name_2) = \frac{sum - cn + 1}{len(name_1) + len(name_2) + sum} \tag{2}$$

Longest Common Substring Comparing Minimum (Maximum) Length. We also take the length of the names into account. Generally speaking, the longer the common substring length of two names, the more similar the two names are. We calculate the Longest Common Substring Comparing Min (Max) Length for users by considering the following relevant formula

$$Sim_{LCSmin}(name_1, name_2) = \frac{len(lcs(name_1, name_2))}{\min(len(name_1), len(name_2))} \tag{3}$$

$$Sim_{LCSmax}(name_1,\ name_2) = \frac{len(lcs(name_1, name_2))}{\max(len(name_1),\ len(name_2))} \tag{4}$$

Normalized Edit Distance. Similar to the Edit Distance, Normalized Edit Distance performs steps add one when adding and deleting, but steps add two when replacing. The specific formula is as follows.

$$Sim_{ned}(name_1,\ name_2) = \frac{len(name_1) + len(name_2) - Sim_{edd}(name_1, name_2)}{len(name_1) + len(name_2)} \tag{5}$$

Jensen-Shannon Divergence of Alphabet Distribution. As mentioJensen-Shannon Divergence of Alphabet Distribution.ned in [15] and [16], the display names "gateman" and "nametag" are very similar because one is the reverse spelling of the other. [17]. The formula for calculating the JS Divergence is as follows.

$$Sim_{JSD} = \frac{1}{2}(KL(p_{name1}|p_{name}) + KL(p_{name2}|p_{name})) \tag{6}$$

$$p_{name} = \frac{1}{2}(p_{name1} + p_{name2}) \tag{7}$$

$$KL(p_{name1}|p_{name}) = \sum_{i=1}^{|p_{name1}|} p_{name1_i}.log\frac{p_{name1_i}}{p_{name_i}} \tag{8}$$

where p_{name} is the alphabet distribution of name, taking David Lee and Dave Wan as an example, the alphabet distribution of both of them is shown below.

Table 3. Alphabet distribution of David Lee and Dave Wan

d		a	v	i	l	e	w	n
$p_{DavidLee}$	2/8	1/8	1/8	1/8	1/8	2/8	p_0	p_0
$p_{DaveWan}$	1/7	2/7	p_0	p_0	p_0	1/7	1/7	1/7

Where p_0 means that the letter does not appear in the string, and we set the position to our preset minimum value of 0.25e–16.

4 Model

This paper is based on users' names to determine whether it is the same person who are using various online social networks, we use a series of supervised machine learning classification models to accomplish this task. The structure of the whole model is as follows (Fig. 2):

4.1 Classifier

For the same dataset, different classification models will have different effects. In this paper, There are ten classification models, including Support Vector Machine (SVM), Logistic Regression (LR), K-Nearest Neighbor (KNN), Gaussian Bayes (GB), Decision Tree (DT), Bagging, Random Forest (RF), Extreme Random Tree (ERT), AdaBoost, and Gradient Boosting Decision Tree (GBDT), are used to classify the data, which are provided by scikit-learn. The result shows that SVM performs best in the three datasets with the accuracy of 92%, 94.5%, 97.1% and AUC values of 95.6%, 95%, 94.4% respectively (Fig. 3, Fig. 4 and Fig. 5).

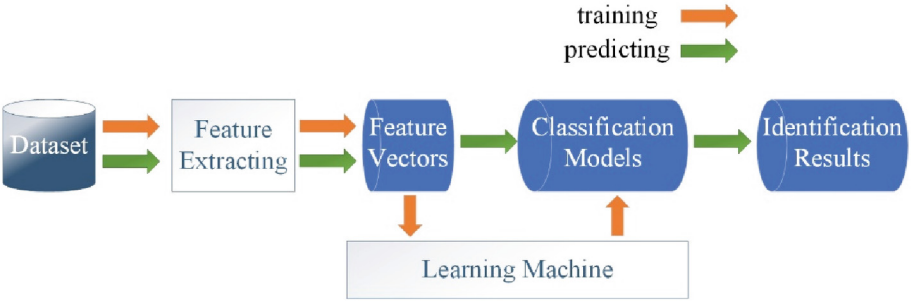

Fig. 2. Identification model procedure

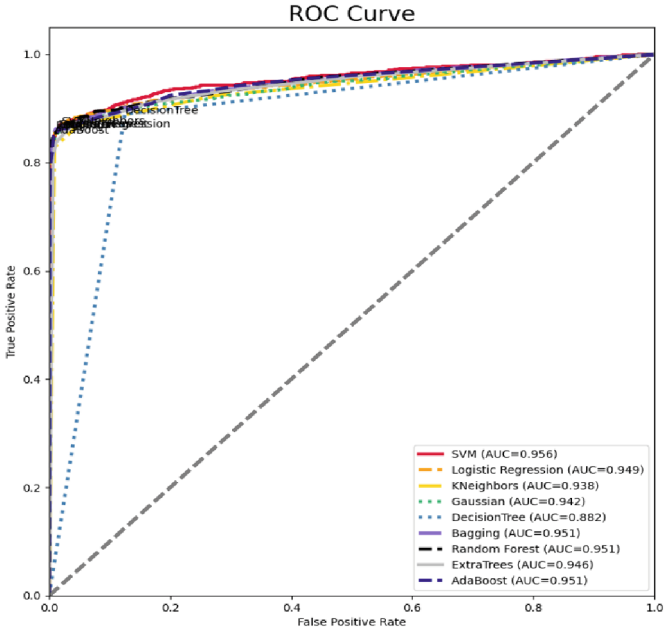

Fig. 3. Identication results of different classfiers in dataset fs-tw

4.2 Evalution Metrics

We introduced the confusion matrix to calculate the goodness of the model. Where a denotes the number of correct predictions in the positive sample, b means the number of wrong predictions in the positive sample, c denotes the number of wrong predictions in the negative sample, and d means the number of correct predictions in the negative sample. Depended on the confusion matrix, our evaluation metrics include ACC, PRE, REC, FNR, and F1 (Table 4).

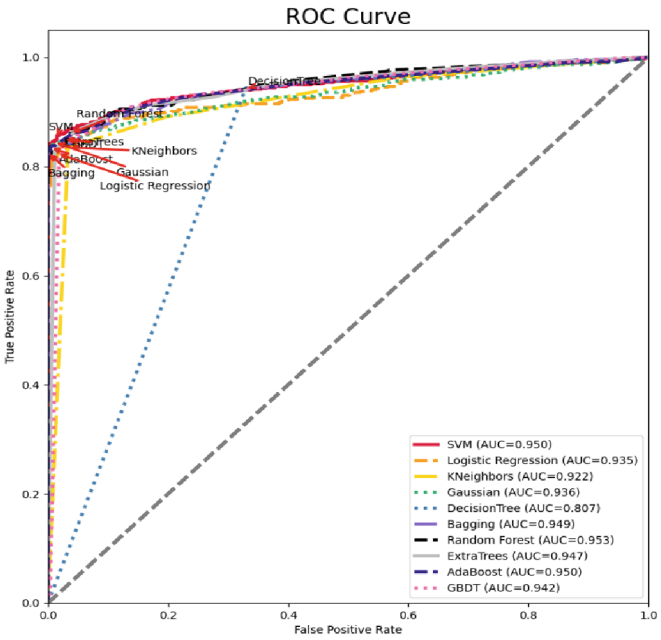

Fig. 4. Identication results of different classfiers in dataset fs-fb

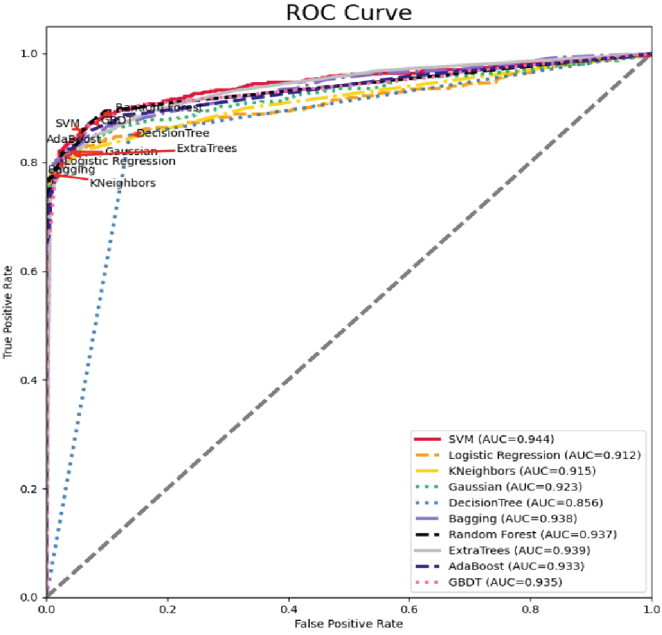

Fig. 5. Identication results of different classfiers in dataset fb-tw

Table 4. Confusion matrix

		Predicted	
		Positive	Negative
True	Positive	a	b
	Negative	c	d

Accuracy (ACC) is the rate of correctly classified samples to the total number of samples, the formula is as follows:

$$ACC = \frac{a+b}{a+b+c+d} \tag{9}$$

Precision (PRE) is the proportion of samples which is predicted to be positive by the model that is also actually positive to the number of samples which is predicted to be positive. The calculation formula is as follows:

$$PRE = \frac{a}{a+c} \tag{10}$$

Recall (REC) is the proportion of the real positive samples to the real positive samples. The formula is as follows:

$$REC = \frac{a}{a+b} \tag{11}$$

F-measure (F1) is the summed average of the precision rate and recall rate and is calculated as follows:

$$F1 = \frac{2 \times PRE \times REC}{PRE + REC} \tag{12}$$

False Negative Rate (FNR) is used to measure the percentage of the minority class that is judged to be wrong as the majority class and is calculated as follows:

$$FNR = \frac{b}{a+b} \tag{13}$$

5 Conclusion

In this paper, we first obtain the names of users from social networking sites, after which we perform feature extraction on the data. And then, the generated feature vector is used as input to the classification model, and the classification model with the best effect is selected according to its performance. The data acquisition cost is low in this paper, which results in a weak classification of users with duplicate names. In future work, we will analyze other attributes of users, including their age, gender, and profile. We will also analyze the user's followership, followings, and the comments made by the user to improve the model's ability to identify users in different data dimension environments.

References

1. Shu, K., Wang, S., Tang, J., et al.: User identity linkage across online social networks: A review. ACM SIGKDD Explorations Newsl **18**(2), 5–17 (2017)
2. Li, H.X., Zhu, H.J., Du, S.G., Liang, X.H., Shen, X.M.: Privacy leakage of location sharing in mobile social networks: Attacks and defense. IEEE Trans. Dependable Secure Comput. **15**(4), 646–660 (2018)
3. Xing, L., Deng, K., Wu, H., et al.: A survey of across social networks user identification. IEEE Access **7**, 137472–137488 (2019)
4. Li, Y., Peng, Y., Zhang, Z., Xu, Q., Yin, H.: Understanding the user display names across social networks. In: Proceedings of International World Wide Web Conference Committee (IW3C2), pp. 1319–1326 (2017)
5. Ma, J., et al.: Balancing user profile and social network structure for anchor link inferring across multiple online social networks. IEEE Access **5**, 12031–12040 (2017)
6. Li, Y., Peng, Y., Zhang, Z., Yin, H., Xu, Q.: Matching user accounts across social networks based on username and display name. World Wide Web **22**(3), 1075–1097 (2018)
7. Ma, J.: Social account linking via weighted bipartite graph matching. Int. J. Commun. Syst. **31**(7) e3471 (2018)
8. Deng, K., Xing, L., Zheng, L., Wu, H., Xie, P., Gao, F.: A user identification algorithm based on user behavior analysis in social networks. IEEE Access **9**, 47114–47123 (2019)
9. K. Deng, L. Xing, M. Zhang, H. Wu, and P. Xie, "A multiuser identification algorithm based on Internet of Things," Wireless Commun. Mobile Comput., vol. 2019, May 2019, Art. no. 6974809
10. Zhao, D., Zheng, N., Xu, M., Yang, X., Xu, J.: An improved user identification method across social networks via tagging behaviors. In: Proceedings of IEEE 30th International Conference on Tools with Artificial Intelligence, pp. 616–622 (Nov 2018)
11. Li, Y., Zhang, Z., Peng, Y., Yin, H., Xu, Q.: Matching user accounts based on user generated content across social networks. Future Gener. Comput. Syst. **83**, 104–115 (2018)
12. Chen, L., Tan, F.: Identity recognition scheme based on user access behavior. In: Proceedings of IEEE 8th Joint International Information Technology and Artificial Intelligence Conference, pp. 125–129 (May 2019)
13. Qi, M., Wang, Z., He, Z., Shao, Z.: User identification across asynchronous mobility trajectories. Sensors **19**(9) (2019), Art. no. 2102
14. Liu, D., Wu, Q., Han, W., Zhou, B.: User identification across multiple websites based on username features. Chin. J. Comput. **38**(10), 2028–2040 (2015)
15. Zafarani, R., Liu, H.: Connecting users across social media sites: A behavioral-modeling approach. In: Proceedings of KDD, pp. 41–49 (2013)
16. Zafarani, R., Tang, L., Liu, H.: User identification across social media. ACM Trans. Knowl. Dis. Data (TKDD) **10**, 1–30 (2015)
17. Li, Y., Peng, Y., Ji, W., Zhang, Z., Quanqing, X.: User identification based on display names across online social networks. IEEE Access **5**, 17342–17353 (2017)