Research Article

Exploiting Two-Level Information Entropy across Social Networks for User Identification

Ling Xing¹, Kaikai Deng², Honghai Wu¹, Ping Xie¹, Mingchuan Zhang¹, and Qingtao Wu¹

¹School of Information Engineering, Henan University of Science and Technology, Luoyang 471023, China

²School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China

Correspondence should be addressed to Ling Xing; xingling_my@163.com

Received 30 April 2021; Revised 8 August 2021; Accepted 2 September 2021; Published 29 September 2021

Academic Editor: Xu Zheng

Copyright © 2021 Ling Xing et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

As the popularity of online social networks has grown, more and more users now hold multiple virtual accounts at the same time. Under these circumstances, identifying multiple social accounts belonging to the same user across different social networks is of great importance for many applications, such as user recommendation, personalized services, and information fusion. In this paper, we mainly aggregate user profile information and user behavior information, then measures and analyzes the attributes contained in these two types of information to implement across social networks user identification. Moreover, as different user attributes have different effects on user identification, this paper therefore proposes a two-level information entropy-based weight assignment method (TIW) to weigh each attribute. Finally, we combine the scoring formula with the bidirectional stable marriage matching algorithm to achieve optimal user account matching and thereby obtain the final matching pairs. Experimental results demonstrate that the proposed two-level information entropy method yields excellent performance in terms of precision rate, recall rate, F -measure ($F1$), and area under curve (AUC).

1. Introduction

A product of the Web 2.0 era, today's social networks provide people with a rich variety of social services. According to research, 42% of users have more than one social network accounts [1]. Since different social networks have their own unique social methods and provide different social services to users, they generate a wealth of social user information; however, each social network account is isolated and there is no direct contact between them. Therefore, the social information generated by user accounts is distributed across multiple social networks. Across social networks, user identification refers to the process of identifying virtual accounts that belong to the same entity user across different social networks. Achieving this type of identification provides assistance for other fields, such as network recommendation and user behavior analysis, in a variety of different ways, and allows the “big data” contained in multisource social networks to be fully exploited.

User identification, also known as user identity resolution and anchor linking, is aimed at identifying the same users across different social networks. The core idea underpinning existing research works involves utilizing user profile information, network topology information, and user behavior information to calculate and analyze whether the user account matching pair represents two instances of accounts created by the same user. The research work based on user profile information mainly focuses on specific attribute information provided by users (e.g., username, birthday, and interests) [2]. Similarity calculation methods are used to calculate the similarities between these attributes used in order to determine whether or not the user account pair is identical. However, existing social networks have begun to focus on user privacy issues, meaning that a portion of attribute information is not easily available. Related works based on network topology information have mainly focused on using a user's friend relationships to identify their accounts across different social networks [3]; however,

these connections between friend relationships are often sparse, meaning that this method encounters certain limitations when applied. Research works based on user behavior information are now primarily aimed at using the content generated by the users to identify said users [4, 5], as user-generated content is easier to access than the other two types of information. Moreover, the content published by users is personalized and can thus be used effectively to map user behavior habits.

User profile information refers to the information that users are required to enter or select when they register for their social accounts. Relevant research has shown that more than 45% of users choose the same display name across different social networks [6]. Therefore, there are a number of largely similar phenomena involved in the process of filling in profile information across different social networks, which provides a reliable basis for user identification. User behavior information mainly refers to the content posted by users on major social networks. Moreover, users often make some of the content they publish public, meaning that this content is easy to access. Intuitively, user behavior information-based user identification could constitute a means of breaking through the limitations of the above two methods.

Accordingly, through the analysis of characteristic information in the process of user identification and a simple comparison of existing work, the contributions of our work are summarized as below:

- (1) We calculate the similarities of the user account found in the user profile information and user behavior information, respectively. Then, we briefly analyze the importance of different attributes in the process of user identification
- (2) To the best of our knowledge, we first propose a two-level information entropy-based weight assignment method to weigh each attribute
- (3) User similarity scoring formula is designed through the similarity calculation and weight assignment of these user attributes
- (4) We use a stable marriage matching algorithm to ensure that the different accounts achieve bidirectional optimality in the matching process in order to further improve user identification performance

In this work, we conduct user identification across two different social networks; for multiple social networks, we can adopt the transitivity law that anchor linkage follows [7] to identify users.

The remainder of this paper is organized as follows. We describe the current related works in Section 2. In Section 3, we formulate the problem of user identification. Section 4 introduces the method of user identification and the details of its implementation, while the experimental results of our proposed method are provided in Section 5. Finally, we summarize this paper and lay out some avenues for future research in Section 6.

2. Related Works

Current studies of user identification across multiple social networks can be divided into three categories, according to the different types of information utilized: user profile information-based matching strategy, network topology information-based matching strategy, and user behavior information-based matching strategy. At present, most studies center around on one of the abovementioned types of information when engaging in user identification, while only a small portion of these studies have fused this information.

2.1. User Profile Information-Based User Identification. Studies engaging in user profile information-based research focus on user profile attributes, which include display name, city, gender, and profile image.

Display names, which are required to be entered by almost all social networks, have been widely studied in an attempt to identify the virtual accounts of users across different social networks. Perito et al. [8] calculated the similarity of display names between accounts and used binary classifiers to identify users. Similarly, Liu et al. [9] focused on determining whether user accounts on different social networks that use the same username belong to the same natural person. Zafarani and Liu [10, 11] proposed a user-mapping method by modeling and analyzing user behavior habits as regards display names. Moreover, the profile image is another attribute that has attracted considerable attention from researchers. Acquisti et al. [12] utilized a face recognition algorithm to enable user identification. Although both display names and profile images can be used to identify users, they are not universal for some large social networks, as many unrelated users may have the same display name; for example, many users have the display name “John Smith” on Facebook.

User identification can be better achieved through the combination of user profile attributes. Iofciu et al. [13] proposed a method that involves measuring the distance between user profile information. Motoyama and Varghese [14] transformed user attributes (gender, education, etc.) into word sets and matched account pairs by calculating the similarities between users. Cortis et al. [15] proposed a user profile identification technique based on weighted ontology to identify users by calculating the similarity between users’ grammatical semantics. Abel et al. [16] aggregated user profiles and implemented identification across systems to match users; a similar study involving multiple social networks can also be found in [17]. Li et al. [18] developed an identification model that combines username and display name, both of which contain rich information redundancy. He and Li [19] designed an identification scheme based on user preferences. They fully exploit the redundant information of the user’s nickname and analyze the topic of the user’s published content. Finally, the above two attributes are combined to identify the user. Mishra [20] developed a pairwise comparison string model that analyzes user profile information among different social networks and adopts different similarity

calculation methods to calculate the similarity of different attributes. Finally, the obtained value is compared with the threshold to generate a final matching pair. Ahmad and Ali [21] proposed a two-step method to examine the problem of matching social accounts across social media services. A good effect is achieved by fusing the username and the network structure of the seed node to identify users. In light of the above, the proposed method reduces the usage of user attributes, while the attributes chosen for use have high accessibility.

2.2. Network Topology Information-Based User Identification. Studies addressing network topology information-based research focus primarily on identifying the same user by utilizing that user's friend relationships. A user's friend relationships are easy to obtain, and malicious imitation and falsification problems are almost impossible. Narayanan and Shmatikov [22] demonstrated that user identification can be accomplished by relying on network topology information; however, the user identification performance achieved in this study required further improvement. Bartunov et al. [23] constructed an objective function by combining profile information and network topology information, then optimized the objective function to obtain optimal account matching pairs. Cui et al. [24] aggregated the similarities between user profile information and graphics to enable mapping from email networks to Facebook networks. Korula et al. [25] converted the user identification problem into a mathematical problem in an attempt to determine whether accounts in different social networks are identical by calculating the similarity of the user graph structure. Tan et al. [26] built models from the users' social relationships, mapped users to a low-dimensional space, and thereby reduced the amount of computation required by the user identification process. Zhou et al. [27] utilized the number of shared seed nodes to measure the similarity between nodes on different social networks and selected the nodes with the highest matching degree for pairing; a similar study was also conducted in [28]. Zhou et al. [29] designed an unsupervised user identification scheme to identify users without the use of seed nodes. This method extracts the friend relationships of each account in the social network as a feature vector, constructs a learning model, and then employs the related similarity calculation method to obtain the similarity between the friend feature vectors. Zhang et al. [30] designed an identification scheme based on graph neural network, which maps network nodes to low-dimensional space and preserves the local and global connection modes, so that the reconstruction social network based on the learning node features is close to the original social network. Recently, Li et al. [31] performed user identification by mining multihop nodes in the user friendship networks. They analyzed the contribution of multihop nodes for user identification and combined the features contained in the display name to achieve better performance and universality of the identification method. However, it should be noted that this information is often sparse, as only a small percentage of users are willing to disclose their friend relationships.

2.3. User Behavior Information-Based User Identification. The primary focus of user behavior information-based user identification studies is the content published by users [32]. Users' behavior information exhibits a certain degree of personalization that can be used to map out the user's behavior habits, making this type of information a good choice for user identification purposes. Narayanan et al. [33] analyzed the writing style of user-published content and calculated the similarity probability of publishing content between different social accounts of the same user. Kong et al. [34] designed a Multi-Network Anchoring (MNA) method for user mapping, which comprehensively calculated the similarity between time-based, space-based, and text content across platforms to achieve user identification. Goga et al. [35] fused the geolocation information, status timestamp, and writing style of published content to identify users. Liu et al. [36] calculated the probability of two accounts belonging to the same natural person with reference to the PoIs, writing style, and locations extracted from user behavior information. Chen et al. [37] proposed a new type of user identification approach by extracting users' spatiotemporal features from user behavior information. Li et al. [38] designed a recognition model based on user behavior information and measured the similarity of user behavior information across the dimensions of space, time, and content; a supervised machine learning algorithm was then used to match the account pairs. Recently, Chen and Tan [39] mapped user access behavior to different time windows, extracting time, text, and sequence features of access behavior to obtain rich semantic content. Finally, a multilayered perception network-based identification model is constructed to identify the user's identity. To the best of our knowledge, many existing works mainly focus on the extraction of user information, while ignoring the analysis of the importance of user information. The state-of-the-art works performed a simple analysis of user information [4, 5]. However, the degree of contribution to distinguishing user information varies greatly, which is not conducive to the balance of weight assignment in the user identification process.

Furthermore, with the development of the Internet of Things [40–45], location information began to enter the researcher's field of vision, which is also a powerful feature for user identification purposes. Hao et al. [46] converted the user trajectory into multiple grid sequences, transformed these sequences into vectors using the TD-IDF model, and then used cosine similarity to calculate the similarity between the vectors obtained. Roedler et al. [47] combined user-generated timestamp information with location information for use in constructing a personalized social behavior pattern designed to address user identification issues. Qi et al. [48] designed an identification scheme based on user trajectory, which analyzes the TOP-N region with the most frequent distribution of user trajectories and uses different similarity calculation methods to measure the similarity between two trajectories. This achieves better identification performance. However, geolocation information is often sparse in social networks, as only a small proportion of users are willing to make their geographic locations public.

3. Problem Definition

This section defines key terms used in the present paper, integrates user profile information and user behavior information to identify the entity users behind multiple social accounts, and elaborates on the process of user identification.

The term “social network” is used to refer to virtual communities and platforms on which people publish, share, and exchange ideas and information. On social networks, people input information within a bounded system that allows them to create a profile and can then communicate with other users [49]. From the above description, it is evident that a social network comprises three important elements: users (with their profile information), behavioral information among users (or content), and friend relationships (or network) [50]. More detailed definitions of these terms are provided below.

Definition 1 (social network). Given a social network $S = \{V, E, C\}$, where V denotes the users, E denotes the user’s profile information, and C denotes the user-generated behavior information.

Delving into the main components of a social network, one can easily determine that both E and C are generated by V . To a certain extent, E and C can therefore be considered as attributes of V , while V is the core element of the social network. Therefore, user identification is of great importance in multiple studies of social networking.

In this study, S_A and S_B are used to represent social networks A and B , respectively.

Definition 2 (entity user). An entity user is a user representing a single unique individual, in combination with his or her profile information, behavior information, and friend relationships.

Definition 3 (user matched pair). Given social networks S_A and S_B , if social accounts V_{Ai} and V_{Bj} belong to the same person in real-life, they then constitute a user matched pair (V_{Ai}, V_{Bj}) .

In this section, we formulate across social networks user identification based on profile information and behavioral information. In Figure 1, we consider two social networks S_A and S_B . Each of these social networks is home to some users whose profile information and behavioral information are presented in Table 1.

Generally speaking, multiple virtual accounts belonging to the same user but created on different social networks tend to be isolated from each other, without any connection between them [34]. The main task of user identification is accordingly to discover the potential relationship between these accounts on different social networks [51]. From Figure 1, in S_A , the user information sets of two people are (E_{Ai}, C_{Ai}) and (E_{Ak}, C_{Ak}) , respectively. Among them, E_{Ai} represents the user’s 17 user profile information, corresponding to T_1^{mp} in Figure 2. C_{Ai} represents the user’s 3

behavioral information, corresponding to T_1^{mb} in Figure 2. (E_{Ak}, C_{Ak}) is the same. Moreover, there is a user whose user information set is (E_{Bj}, C_{Bj}) in S_B . The following two problems need to be solved for user identification to be achieved:

- (1) There are two accounts V_{Ak} and V_{Bj} ; can we determine whether these two accounts belong to the same person?
- (2) There are two accounts V_{Ai} and V_{Bj} ; can we determine whether these two accounts belong to two different persons?

If two user information sets from different social networks are sufficiently similar, they are likely to belong to the same person. The higher their attribute similarity, moreover, the higher the probability that they are the same person. User identification issues can thus be defined as follows:

$$f(V_{Ai}, V_{Bj}) = g(\text{Score}_{ij}) = \begin{cases} 1 & \text{if } V_{Ai} \text{ and } V_{Bj} \text{ belong to} \\ & \text{the same person;} \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

where Score_{ij} denotes the similarity score of V_{Ai} and V_{Bj} .

It should be noted here that, for various reasons, some people have multiple accounts in the same social network; however, we usually consider these accounts to be independent and belonging to different persons, which is to say that we only identify one of the accounts.

4. Proposed Method

4.1. User Identification Framework. In this paper, we mainly aim to combine users’ profile information and behavior information to identify the entity users behind multiple social accounts. We fuse these two types of information for three main reasons.

- (1) User profile information and user behavior information are easier to obtain than user network topology. In addition, these two types of user information can more intuitively reflect the user’s living habits and style on different social networks
- (2) User profile information and user behavior information have similarity across social networks
- (3) User topology will have the contribution of single-hop nodes and multihop nodes. If the user topological structure is weighted directly, it will cause the problem of weight imbalance among the user nodes

Based on the above analysis, we focus on integrating user profile information and user behavior information. Through in-depth mining of the two types of user information, a two-level information entropy-based weight assignment algorithm is proposed to solve the problem of weight imbalance caused by multiattribute information.

FIGURE 1: Cross-network research to merge various SNs.

TABLE 1: User information used for user identification across social networks.

Attributes	Information type	Attributes	Information type
Username	Profile information	Political views	Profile information
Gender	Profile information	Favorite quotes	Profile information
Birthday	Profile information	Relationship status	Profile information
Hometown	Profile information	Websites	Profile information
Biography	Profile information	Emails	Profile information
URL	Profile information	Education	Profile information
Interests	Profile information	Professional experience	Profile information
Current city	Profile information	Frequent mode	Behavioral information
Languages	Profile information	Punctuation mark prefer	Behavioral information
Religion views	Profile information	Status timestamp	Behavioral information

As shown in Figure 2, given three user accounts v_1^n , v_1^m , and v_2^k form two different social networks S_A and S_B , respectively. Where each user account contains 17 user profile information (such as T_1^{mp}) and 3 user behavior information (such as T_1^{mb}). Let us select accounts v_1^n and v_2^k as an example. This pair of accounts is calculated to obtain the similarity S_{nk} ; similarly, S_{mk} denotes the similarity calculation of 20 user attribute values between users m and k . Then, a weight assignment method based on two-level information entropy needs to be used to weight each user information, and a similarity vector X_{nk} based on weight is obtained. Finally, the bidirectional stable marriage matching algorithm is used for account matching. If $\gamma = 1$, the account matches; otherwise, the two accounts belong to different persons.

4.2. User Profile Information Analysis. In Section 3, we divide the user profile information into different dimensions according to different feature attributes. The similarity calculation methods are used to calculate the similarity of each dimension of the two accounts across different social networks, after which the similarity values are compared with their corresponding thresholds in order to obtain the results of the user profile information comparison across the different dimensions.

Before the similarity calculation is performed, the profile information must be generalized to ensure that the information is unified across the different social networks. For example, the user's gender may exist in different social networks as "boy, girl," "man, woman," etc., and this paper uniformly generalizes to "boy, girl." In addition, since the types of user information we analyze are different, different similarity calculation methods are needed. In this paper, there are three key methods adopted to perform the similarity calculation of profile information with different attributes: Dice coefficient (it is one of the common methods to evaluate the classification effect, and the same can be used as a loss function to measure the loss between the classification result and the label), cosine similarity (this method mainly takes the user's preference as a point in the n -dimensional coordinate system and forms a straight line (vector) by connecting this point with the origin of the coordinate system. The similarity value between two users is the cosine value of the angle between the two vectors), and exact matching (specific user information, such as gender). These methods are explained in more detail below:

- (1) *Dice Coefficient.* Let a and b represent two sets of strings, respectively. The numerator denotes the length of the intersection of the two sets, while the denominator denotes the sum of the lengths of the

FIGURE 2: Across social networks user identification implementation framework.

two sets. For example, for the strings $a = \{\text{play, music, basketball}\}$ and $b = \{\text{music, reading, football}\}$, the intersection information is $\{\text{music}\}$, while $|a| = 3$ and $|b| = 3$, meaning that the Dice coefficient of these two strings is $2[1/(3 + 3)] \approx 0.33$

$$S_{\text{Dice}}(a, b) = 2 \frac{|a \cap b|}{|a| + |b|}. \quad (2)$$

- (2) *Cosine Similarity*. This method is often used to calculate the similarity between two vectors. We quantize two strings into word vectors; here, let x_i and y_i denote the word vectors, while i denotes the i th dimension of the word vector, and m denotes the word vector dimension. Therefore, the cosine similarity of the two strings is as follows:

$$S_{\cos}(x, y) = \frac{\sum_{i=1}^m (x_i \times y_i)}{\sqrt{\sum_{i=1}^m (x_i)^2} \times \sqrt{\sum_{i=1}^m (y_i)^2}}. \quad (3)$$

- (3) *Exact Matching*. Under this method, the attribute information of the two accounts must be identical. For example, in the matching process, the gender of two accounts must match exactly (e.g., they are either male or female).

4.3. User Behavior Information Analysis. Generally speaking, user behavior information uses the content published by the user to deal with the similarity between social accounts. Here, content published by the user refers primarily to the text data and the time state accompanying the text data; that is, the dynamic time generated by the blog post and the blog post published by the user. Although user

profile information is important to user identity matching, there will be some information loss problems caused by privacy protection, meaning that the performance of any identification based solely on user profile information will not be perfect. Accordingly, in this paper, we analyze user behavior information on the basis of user profile information. In short, user identification can be better achieved by fusing the above two types of information.

4.3.1. User Blog Data Similarity Calculation. This paper draws on some of the ideas underpinning the concept of frequent pattern mining to propose a method for calculating the similarity of user blog data. Some mining algorithms can be used to determine potential connections between users' behaviors. The specific details of the frequent pattern mining process are explained in Figure 3 below.

Step 1. Word segmentation is performed on each user's blog. As can be seen from Figure 3, the user's blog post constitutes the TID, which is the transaction ID. Each blog post contains a different number of frequent items, which is the set $\{I1 - I5\}$. Among them, the set formed by the blog post is the word segmentation performed by the frequent pattern mining algorithm. After word segmentation is complete, the blog post is treated as consisting of many distinct words, each of which forms a transaction T_i ; as such, all of a user's blog posts form the transaction set $T = [T_1, T_2, \dots, T_i]$.

Step 2. The first iteration of the algorithm; traversing all the items in the transaction set T and calculating its support constitutes one-item set C_1 . In this paper, the minimum support for the one-item set is set to two. With reference to the support requirement, item sets that do not satisfy the condition are filtered out until only the frequent one-item set L_1 remains.

Step 3. L_1 is connected with itself to obtain the two-item set C_2 . For example, $\{I1\}$ and $\{I2\}$ in L_1 are connectable, and the item set $\{I1, I2\}$ can be obtained after the connection. By analogy, we connect the connectable item set to get the candidate set C_2 and obtain the support count corresponding to the related item set. Moreover, we set the support of C_2 is one. The transaction set T is scanned to filter frequent items that do not satisfy the support, with the result that the frequent two-item set L_2 is obtained.

Step 4. Similarly, the frequent three-item set L_3 , frequent four-item set L_4 , ..., frequent n -item set L_n are generated in sequence until all generated item sets C are unable to satisfy the minimum support; at this point, the algorithm ends. The example presented in Figure 3 ends with the frequent four-item set L_4 .

After the above process is complete, the frequent items and their support counts have been obtained. Accordingly, the similarity calculation formula of the blog data of users i and j is as follows:

$$S_{ij} = \sum_{F_n \in i \cap j} [(1 + C_{Ai}(F_n)) \times (1 + C_{Bj}(F_n))]^{C_{F_n}}, \quad (4)$$

where F_n denotes the frequent items shared by users i and j , $C_{Ai}(F_n)$ and $C_{Bj}(F_n)$ denote the support counts of frequent items F_n of users i and j , respectively, while C_{F_n} denotes the number of item set of F_n . The "1" in the formula is added to ensure that high-frequency items are avoided.

4.3.2. Punctuation Marks Similarity Calculation. Punctuation marks, when used in the writing of blog posts, can clearly reflect the user's personal habits of behavior where writing is concerned. Therefore, this information can also be measured and analyzed as an attribute for use in determining user's identity. In this subsection, a punctuation mark vector is generated to store the frequency at which punctuation marks appear in the users' blog posts. The measure of punctuation similarity is aimed at determining whether users utilize similar punctuation marks when writing a blog post. In Table 2, we present a case in which two users, Ben and Emily, use different punctuation marks in the blog posts they publish.

In order to calculate the similarity between the punctuation marks used by each user account, each user's punctuation marks are quantized into a vector; each element of the vector is $P_i = c_i/n$. Here, c_i is the count of each punctuation mark, and n is the total number of blog posts. In this way, we can obtain the user account's punctuation mark vector. In this paper, we calculate the similarity between different account punctuation vectors via cosine similarity. The specific formula is shown in (3).

4.3.3. State Timestamp Similarity Calculation. The same user has a high probability to generate consistent dynamic time across different social networks; this attribute can also fully reflect the user's behavior habits. State timestamp similarity is a measure of a user's behavior characteristics according to the dynamic numbers (the number of posts published by users at each time interval) generated by the user in different time periods, thereby enabling the similarity between the two user accounts to be more accurately calculated.

We first need to tag the time at which the user posted each status. The dynamic number of each time period is counted and divided by the total number of blog posts, from which the average dynamic number of each time period is obtained; this average dynamic number of each time period can constitute a 24-dimensional user timestamp vector. Finally, formula (5) is used to calculate the similarity of the status timestamps between different accounts.

$$S_t(i, j) = \sum_{t=1}^T |v_{it} - v_{jt}|. \quad (5)$$

Here, v_{it} and v_{jt} denote the average dynamic number of users i and j over period t , while T denotes the dimension of the timestamp.

4.4. Two-Level Information Entropy-Based Weight Assignment Algorithm

4.4.1. User Attribute Weight Assignment Analysis. Since different user attributes have different effects on user

FIGURE 3: An example of frequent pattern mining.

TABLE 2: The distribution of punctuation marks in user blog posts.

	Period	Comma	Exclamation	Colon	Swung dash	Hyphen	Apostrophe	n
Ben	112	168	12	3	0	12	36	100
Emily	10	36	3	2	27	0	35	100

identification performance, we need to assign reasonable weights for each of the attribute items. Following earlier works [4, 5], Figure 4 illustrates the contribution of a single attribute used for user identification.

It can be clearly seen from Figure 4(a) that the username has different distributions of similarity between the same user and different user when performing account matching; therefore, it can be concluded that this attribute item has strong discriminability, and thus that the weight assigned to it should be large. Moreover, it can be clearly seen from Figure 4(b) that when account matching is performed using the interest attribute, the user distribution is relatively uniform, meaning that it is not easy to determine whether or not users are identical; hence, the weight assigned to this attribute should be relatively small. In summary, there are differences in the contribution made by each attribute to user identification; it is therefore vital to assign appropriate weights to each attribute.

4.4.2. Weight Assignment Algorithm. After calculating the similarity between user attributes, we need to assign corresponding weights to each attribute to further improve user identification performance. Both traditional expert subjective weighting methods and objective weighting methods have their limitations, which include poor robustness, dependence on large amounts of sample data, and poor generality.

In order to solve the above problems associated with the weight assignment scheme, the present paper proposes a two-level information entropy-based weight assignment method. The basic idea behind the entropy weight method is that the greater the difference in the index, the greater the difference in weights; therefore, we use the concept of information entropy to solve the problem of weight assignment in user identification. According to the definition of information entropy, for any random variable x , the formula is as follows:

$$E(x) = - \sum_{x \in X} p(x) \log p(x), \quad (6)$$

where $p(x)$ denotes the value probability of this attribute.

Based on the information entropy concept, the posterior probability of each attribute is further calculated, allowing the impact of each attribute on user identification performance to be more accurately determined. The one-level weight assignment of the user attribute can be obtained by combining the posterior probability and the information entropy, as follows:

$$W(x) = -p(y_s | s) \sum_{x \in X} p(x) \log(p(x)), \quad (7)$$

FIGURE 4: The impact of user attributes on user identification performance.

where $p(y_s | s)$ is the posterior probability of the attribute, i.e., the probability that the same user attribute y_s is consistent, x is an attribute, and X is all attributes. The posterior probability and $p(x)$ can be obtained via statistical means.

Softmax is an extremely common and important function in machine learning and deep learning applications, particularly in multiclass scenarios. It represents some input mappings as real numbers between zero and one; moreover, normalization guarantees that the sum will be one, meaning that the sum of the probabilities of multiple classifications is also exactly one. The output of Softmax characterizes the relative probability between different categories. Accordingly, we use the Softmax concept to assign a two-level weight to the user's attributes. After obtaining the one-level weight of the user attribute, the weight values of all attributes are combined into an array $Z = (z_1, z_2, \dots, z_n)$, which is used as the Softmax input.

As shown in Figure 5, the one-level weight value of each attribute is calculated using the Softmax function, after which the probability corresponding to each attribute is output. Using Softmax should address the problem of numerical overflow. Due to the exponential operation, if the Z value is large, the value after the exponential operation is often exposed to the possibility of overflow, meaning that some numerical processing for Z is required. This paper therefore multiplies the elements in Z by one-tenth. As mentioned above, the probability formula for user attributes is calculated as follows:

$$P_i = \frac{e^{\theta z_i}}{\sum_{j=1}^n e^{\theta z_j}} \text{ and } 0 < P_i < 1, \sum_i P_i = 1, \theta = \frac{1}{10}. \quad (8)$$

After obtaining the probability of each attribute, the attribute entropy value can be redefined as:

$$E(P_i) = -\frac{e^{\theta z_i}}{\sum_{j=1}^n e^{\theta z_j}} \log \frac{e^{\theta z_i}}{\sum_{j=1}^n e^{\theta z_j}}. \quad (9)$$

Moreover, since the entropy value is inversely proportional to the weight, the variant entropy value R_i can be expressed as:

$$R_i = \frac{1}{E(P_i)}. \quad (10)$$

Finally, the two-level weight of each attribute is

$$W'_i = \frac{\sum_{j=1}^n R_j}{R_i}. \quad (11)$$

As shown in Figure 6, we compare the variances (y -axis) of three methods of weight assignment: empirical probability-based weights, posterior probability-based weights, and two-level information entropy-based weights. The specific values obtained by these three weight assignment methods are 0.0031, 65.9918, and 285.8982, respectively. The experimental data also fully demonstrates the effectiveness of the proposed method. Note that the blue histogram corresponds to the left axis; the red histogram corresponds to the right axis.

4.5. User Account Matching

4.5.1. Scoring Formula. The user similarity score formula can be constructed by calculating the similarity between the user attributes and assigning weights accordingly. In order to improve the efficiency of similarity calculation between user accounts, this paper uses a stable marriage matching algorithm [18] to match account pairs across social networks; moreover, it also utilizes a bidirectional matching strategy

FIGURE 5: Softmax workflow.

FIGURE 6: Variance comparison of user attribute weight assignments.

to further improve user identification performance. The user similarity score formula is as follows:

$$\text{Score}_{ij} = \sum_{x=1}^u W'_x S_{ij}^x. \quad (12)$$

Here, Score_{ij} denotes the final matching score, W'_x represents the weight of the x th attribute of the account, while S_{ij}^x denotes the similarity of the x th attribute of the two accounts i and j , and the value of u is twenty. The size of the Score_{ij} is used to determine whether the entity users behind the two social accounts are identical.

4.5.2. Bidirectional Stable Marriage Matching Algorithm. This step involves constructing the scoring formula between user accounts, then using the bidirectional stable marriage matching algorithm to achieve user account matching. Generally speaking, the best user identification is to satisfy one-to-one matching. The core idea of the TIW-UI algorithm we

proposed is still stable marriage matching, so the corresponding algorithm complexity is $o(n^2)$, which is the same as state-of-the-art matching algorithm complexity [5, 29, 52]. It is worth noting that we mainly focus on the weight assignment of user information. The following is the specific process of the matching algorithm used in this paper.

Step 1. Matching each user account in social network S_A with all user accounts in social network S_B via the scoring formula. The specific process is as outlined in Algorithm 1.

Step 2. Each account in S_A is matched with the top-ranked account in S_B in descending order of final score. If the account in S_B has not been matched with any other account in S_A , then the account is matched with the current account in S_A . If, however, the account has also been matched with other accounts in S_A , then all accounts matching the account need to be scored, until the account with the highest score is finally selected as the matching pair.

```

input:  $S_A, S_B$ .
Output: Candidate matching pair sets.
1: Candidates  $\leftarrow \{\}$ 
   For each  $V \in S_A$  do
      $V_{top-n} \leftarrow$  select account from  $S_B$ 
      $\vec{t} \leftarrow \{(V, V_{top-1}), (V, V_{top-2}), \dots (V, V_{top-n})\}$ 
     Labels, Scores  $\leftarrow$  classification( $\vec{t}$ )
      $V_{candidate} \leftarrow V_{top-n}, (\text{labels}, \text{scores})$ 
     Candidates  $\leftarrow \{V, V_{candidate}\}$ 
   End
   Return candidates

```

ALGORITHM 1: Select candidate account matching pairs.

```

input:  $S_A, S_B$ , candidates.
Output: Final matching pairs  $F_{match}$ .
1: For each account  $V_{Ai} \in S_A$ , the score results is
   Named  $\{V_{top-n}^A\}$ 
   For each account  $V_{Bj} \in S_B$ , the score results is
   Named  $\{V_{top-n}^B\}$ 
   Select account  $V_{Bj}$  from  $\{V_{top-n}^A\}$ 
   If  $V_{Bj}$  not matched
     Add  $(V_{Ai}, V_{Bj})$  to  $F_{match}$ 
     Set  $V_{Ai}$  and  $V_{Bj}$  as matched
   Else
     Compare the scores of  $V_{Ai}$  and  $V_{Ak}$  in
      $\{V_{top-n}^B\}$  (assume  $V_{Bj}$  has been marked with
      $V_{Ai}$ )

```

ALGORITHM 2: Bidirectional stable marriage matching.

Step 3. If there are still accounts that do not match, return to Step 2. The overall detailed bidirectional matching of the algorithm is outlined in Algorithm 2.

5. Analysis of Experimental Results

Several experiments were conducted to more intuitively illustrate the effectiveness of the proposed method. All experiments were performed using a computer with 8 G RAM and 2.4 GHz CPU. User profile information and user behavior information from two major social networks, namely, Facebook and Twitter, were selected for user identification across social networks. [53] provides an open dataset that includes five foreign mainstream social networks. The ratio between the training set and the test set is set as 3:1.

In order to facilitate the comparison between the proposed algorithm and other algorithms, this paper uses the most widely used user identification metrics, namely, precision rate, recall rate, $F1$, and AUC. The relevant formulae are as follows:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad (13)$$

$$\text{Recall} = \frac{TP}{TP + FN}, \quad (14)$$

$$F1 = 2 \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}. \quad (15)$$

AUC refers to the area under the ROC curve. The false positive rate (FPR) is defined as the x -axis, while the true positive rate (TPR) is defined as the y -axis. Since the results of this paper are divided into two categories—i.e., same entity user and different entity users—AUC can also be used to measure identification results.

$$\text{TPR} = \frac{TP}{TP + FN}, \quad (16)$$

$$\text{FPR} = \frac{FP}{FP + TN}, \quad (17)$$

where TP denotes the case in which the two accounts are the same user and the matching result indicates the same user, TN denotes the two accounts not being the same user while the matching result indicates different users, FP denotes the two accounts belonging to different users while the matching result indicates the same user, and FN denotes an instance in which the two accounts belong to the same user, but the matching result indicates that the users are different.

5.1. The Impact of Weight on User Identification Performance. The two-level information entropy-based weight assignment algorithm is used to assign weights to user attributes with the aim of improving identification performance. In order to verify the algorithm's effectiveness, we use the control variable method to analyze the experimental results of three different weight assignment methods.

- (1) *The Empirical Probability-Based Weight Assignment Method (EW).* There are two ways to obtain weights, one of which is calculated by using user history information. Second, when historical information is not available or the data is incomplete, it can be obtained based on expert experience [4]
- (2) *The Posteriori Probability-Based Weight Assignment Method (PW).* The user information has been obtained, and the probability that the user information is the same among the same attributes on different social networks is judged [5]

From the description of the above two methods, it can be seen that user information cannot be analyzed in a fine-grained manner. Only relying on the initially obtained data probability for weight assignment will cause category tilt and make the weight assignment unbalanced. Therefore, we propose a two-level information entropy-based weight assignment method (TIW). As shown in Figure 7 and Table 3, we can draw the following conclusions:

FIGURE 7: Comparison of evaluation metrics for three weight distribution methods.

TABLE 3: Comparison of EW, PW, and TIW evaluation metrics.

Number of users	EW (error range ± 0.001)				PW (error range ± 0.001)				TIW (error range ± 0.001)			
	Precision	Recall	F1	AUC	Precision	Recall	F1	AUC	Precision	Recall	F1	AUC
1000	0.859	0.850	0.854	0.858	0.961	0.860	0.908	0.960	0.976	0.868	0.919	0.965
2000	0.835	0.845	0.840	0.849	0.954	0.852	0.900	0.935	0.964	0.861	0.910	0.960
3000	0.824	0.836	0.830	0.827	0.949	0.847	0.895	0.945	0.961	0.855	0.905	0.952
4000	0.815	0.804	0.809	0.804	0.937	0.835	0.883	0.941	0.946	0.846	0.893	0.948
5000	0.794	0.795	0.794	0.755	0.930	0.830	0.877	0.925	0.938	0.840	0.886	0.934
6000	0.761	0.770	0.765	0.731	0.910	0.811	0.858	0.906	0.927	0.826	0.874	0.909

- (1) Following analysis of the experimental results, we can clearly see that TIW performs better than EW and PW
- (2) EW, PW, and TIW all exhibited a downward trend in terms of precision rate, recall rate, $F1$, and AUC. This is because when the total number of social accounts is low, the possibility of higher similarity between accounts is relatively low; as the number of user accounts increases, there will be cases in which similarity between the accounts is high, but the entity users are not the same. When the above situation occurs, this will have a negative impact on the final matching results
- (3) Although EW, PW, and TIW all show a downward trend in precision rate, recall rate, $F1$, and AUC, it can be seen from the figure that the decline rate of TIW is small, while the decline rate of the other two schemes is relatively large
- (4) Through comparative analysis of the experiment results, the effectiveness of the proposed algorithm can be intuitively grasped. Overall, TIW greatly

FIGURE 8: Performance comparison of user account matching algorithms.

TABLE 4: Evaluation metrics of user account matching algorithms.

Evaluation metrics (error range)	RCM	RFCA-SMM	TIW-UI
Precision (± 0.001)	0.935	0.961	0.972
Recall (± 0.001)	0.881	0.862	0.873
F1 (± 0.001)	0.907	0.908	0.918
AUC (± 0.001)	0.910	0.960	0.967

improves the precision rate, recall rate, F1, and AUC of user identification across social networks

5.2. The Impact of Different Matching Algorithms on User Identification Performance. This paper proposes a two-level information entropy-based across social network user identification algorithm (TIW-UI). The effectiveness of the proposed algorithm is illustrated through comparison with the random forest confirmation algorithm, based on stable marriage matching (RFCA-SMM) [5], and the ranking-based cross-matching algorithm (RCM) [52]. RFCA-SMM analyzes user-generated data and combines posterior probability and information entropy to assign weight to user data. However, the weight assignment method is not able to fully reflect the differences between user attributes. The identification performance of the RCM is largely influenced by the number of seed users, i.e., the number of known matching pairs. In cases where it is not possible to know in advance which accounts belong to the same entity user across two social networks, the RCM requires improvement in terms of its identification performance. Since the algorithm for user identification in this paper is untagged, the experimental results of the three algorithms are analyzed using an untagged dataset.

From Figure 8 and Table 4, it can be seen that the proposed TIW-UI is superior to RFCA-SMM and RCM in terms of precision rate, recall rate, F1, and AUC. TIW-UI analyzes a total of 17 user profile information attributes and 3 user behavior information attributes and thereby comprehensively calculates the similarity in users' information across social networks. The two-level weight assignment is

adopted to the attributes used. Compared with RFCA-SMM and RCM, different attributes make substantially different contributions to the identification results, which improve the overall user identification performance. Moreover, TIW-UI and RFCA-SMM achieve some improvement compared with RCM in terms of the evaluation metrics, mainly because these two algorithms do not take the social network structure into account [52]. In order to more clearly illustrate the effectiveness of the proposed algorithm. As shown in Table 4, we can clearly see the performance comparison of the three algorithms among different evaluation metrics. In terms of precision rate, TIW-UI improved by 3.7% and 1.1% compared to RCM and RFCA-SMM. In terms of recall rate, TIW-UI is 1.1% higher than RFCA-SMM, but slightly lower than RCM. This is because RCM uses cross matching to filter out the wrong matching pair to some extent. TIW-UI increased by 1.1% and 1% in terms of F1 compared to RCM and RFCA-SMM, and in terms of AUC, it increased by 5.5% and 0.7% compared to RCM and RFCA-SMM. In general, the proposed algorithm achieves better user identification performance than the other two algorithms.

6. Conclusions

This study addressed the problem of across social networks user identification and provided an innovative solution. As the most relevant types of user information, user profile information, and user behavior information are key to identifying the unique users behind different social network accounts. In this paper, the similarity between the above two types of user information is calculated and analyzed, and a two-level information entropy-based weight assignment algorithm is employed to weight each user attribute appropriately. Finally, we combine the scoring formula with the bidirectional stable marriage matching algorithm to obtain the optimal account matching results.

In future research works, user privacy protection will also become an urgent problem to be solved. There should be a complementary relationship between user identification and user privacy protection. The problem of user privacy leakage caused by user identification should be minimized in the process of research. We will aim to utilize the idea of game theory to achieve a balance between user identification and user privacy protection. If this idea can be translated into a workable form, it will solve the data constraint problem affecting most existing identification algorithms. Moreover, the protection of user privacy data is also effectively realized.

Data Availability

Data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (Grant 62071170, Grant 61771185, Grant 62072158, and Grant 61772175) and the Program for Innovative Research Team in University of Henan Province (21IRTSTHN015).

References

- [1] Y. T. Zhang, J. Tang, Z. L. 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, New York, NY, USA, 2015.
- [2] Y. J. Li, Y. Peng, W. L. Ji, Z. Zhang, and Q. Q. Xu, "User identification based on display names across online social networks," *IEEE Access*, vol. 5, pp. 17342–17353, 2017.
- [3] K. Shu, S. H. Wang, J. Tang, R. Zafarani, and H. Liu, "User identity linkage across online social networks," *ACM SIGKDD Explorations Newsletter*, vol. 18, no. 2, pp. 5–17, 2016.
- [4] K. Deng, L. Xing, L. Zheng, H. Wu, P. Xie, and F. Gao, "A user identification algorithm based on user behavior analysis in social networks," *IEEE Access*, vol. 7, pp. 47114–47123, 2019.
- [5] K. Deng, L. Xing, M. Zhang, H. H. Wu, and P. Xie, "A multi-user identification algorithm based on internet of things," *Wireless Communications and Mobile Computing*, vol. 2019, Article ID 6974809, 11 pages, 2019.
- [6] 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.
- [7] J. Zhang and S. Y. Philip, "Multiple anonymized social networks alignment," in *2015 IEEE International Conference on Data Mining*, pp. 599–608, Atlantic City, NJ, USA, 2015.
- [8] D. Perito, C. Castelluccia, M. A. Kaafar, and P. Manils, "How unique and traceable are usernames?," *Privacy Enhancing Technologies*, vol. 6794, pp. 1–17, 2011.
- [9] J. Liu, F. Zhang, X. Song, Y. Song, C. Lin, and H. Hon, "What's in a name?: an unsupervised approach to link users across communities," in *Proceedings of the sixth ACM international conference on Web search and data mining*, pp. 495–504, Rome, Italy, 2013.
- [10] R. Zafarani and H. Liu, "Connecting corresponding identities across communities," in *Third International AAAI Conference on Weblogs and Social Media*, vol. 6794, pp. 354–357, San Jose, California, USA, 2009.
- [11] 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, Illinois, USA, 2013.
- [12] A. Acquisti, R. Gross, and F. Stutzman, *Privacy in the Age of Augmented Reality*, Proceedings of the National Academy of Sciences, 2011.
- [13] T. Iofciu, P. Fankhauser, F. Abel, and K. Bischoff, "Identifying users across social tagging systems," in *Fifth International AAAI Conference on Weblogs and Social Media*, pp. 522–525, Barcelona, Catalonia, Spain, 2011.
- [14] M. Motoyama and G. Varghese, "I seek you: searching and matching individuals in social networks," in *Proceedings of the eleventh international workshop on Web information and data management*, pp. 67–75, Hong Kong, China, 2009.
- [15] K. Cortis, S. Scerri, I. Rivera, and S. Handschuh, "An ontology-based technique for online profile resolution," in *International Conference on Social Informatics*, vol. 8238, pp. 284–298, Kyoto, Japan, 2013.
- [16] F. Abel, E. Herder, G. J. Houben, N. Henze, and D. Krause, "Cross-system user modeling and personalization on the social web," *User Modeling and User-Adapted Interaction*, vol. 23, no. 2-3, pp. 169–209, 2013.
- [17] X. Mu, F. Zhu, E.-P. Lim, J. Xiao, J. Wang, and Z.-H. Zhou, "User identity linkage by latent user space modelling," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1775–1784, San Francisco, CA, USA, 2016.
- [18] 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.
- [19] Z. He and W. Li, "Research on user identification across multiple social networks based on preference," in *2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)*, pp. 122–128, Nanjing, China, 2018.
- [20] R. Mishra, "Entity resolution in online multiple social networks (@Facebook and LinkedIn)," *Emerging Technologies in Data Mining and Information Security*, vol. 813, pp. 221–237, 2019.
- [21] W. Ahmad and R. Ali, "User identification across multiple online social networks using cross link attribute and network relationship," *Journal of Interdisciplinary Mathematics*, vol. 23, no. 1, pp. 205–214, 2020.
- [22] A. Narayanan and V. Shmatikov, "De-anonymizing social networks," in *2009 30th IEEE Symposium on Security and Privacy*, pp. 173–187, Oakland, CA, USA, 2009.
- [23] S. Bartunov, A. Korshunov, S. Park, W. Ryu, and H. Lee, "Joint link-attribute user identity resolution in online social networks," in *Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining, Workshop on Social Network Mining and Analysis*, pp. 1–9, Beijing, China, 2012.
- [24] Y. Cui, J. Pei, G. T. Tang, W. S. Luk, D. X. Jiang, and M. Hua, "Finding email correspondents in online social networks," *World Wide Web*, vol. 16, no. 2, pp. 195–218, 2013.
- [25] N. Korula and S. Lattanzi, "An efficient reconciliation algorithm for social networks," *Proceedings of the VLDB Endowment*, vol. 7, no. 5, pp. 377–388, 2013.
- [26] S. Tan, Z. Guan, D. Cai, X. Qin, J. Bu, and C. Chen, "Mapping users across networks by manifold alignment on hypergraph," in *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, vol. 14, pp. 159–165, Quebec City, QC, Canada, 2014.
- [27] X. Zhou, X. Liang, H. Zhang, and Y. Ma, "Cross-platform identification of anonymous identical users in multiple social media networks," *Transactions On Knowledge And Data Engineering*, vol. 28, no. 2, pp. 411–424, 2016.
- [28] Y. Qu, S. Yu, W. Zhou, and J. Niu, "FBI: friendship learning-based user identification in multiple social networks," in *2018 IEEE Global Communications Conference*, pp. 1–6, Abu Dhabi, United Arab Emirates, 2018.
- [29] X. Zhou, X. Liang, X. Du, and J. Zhao, "Structure based user identification across social networks," *IEEE Transactions on Knowledge & Data Engineering*, vol. 30, no. 6, pp. 1178–1191, 2018.

- [30] W. Zhang, K. Shu, H. Liu, and Y. Wang, "Graph neural networks for user identity linkage," 2019, <https://arxiv.org/abs/1903.02174>.
- [31] 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.
- [32] L. Xing, K. Deng, H. Wu, P. Xie, and J. Gao, "Behavioral habits-based user identification across social networks," *Symmetry*, vol. 11, no. 9, article 1134, 2019.
- [33] A. Narayanan, H. Paskov, N. Z. Gong et al., "On the feasibility of internet-scale author identification," in *2012 IEEE Symposium on Security and Privacy*, pp. 300–314, San Francisco, CA, USA, 2012.
- [34] 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 & Knowledge Management*, pp. 179–188, San Francisco, CA, USA, 2013.
- [35] 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, Rio de Janeiro, Brazil, 2013.
- [36] 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.
- [37] 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*, pp. 517–526, Singapore, 2017.
- [38] 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.
- [39] L. Chen and F. Tan, "Identity recognition scheme based on user access behavior," in *Identity recognition scheme based on user access behavior*, pp. 125–129, Chongqing, China, 2019.
- [40] L. Liu, Y. Song, H. Zhang, H. Ma, and A. V. Vasilakos, "Phy-sarum optimization: a biology-inspired algorithm for the steiner tree problem in networks," *IEEE Transactions on Computers*, vol. 64, no. 3, pp. 818–831, 2015.
- [41] M. Zhang, M. Yang, Q. Wu, R. Zheng, and J. Zhu, "Smart perception and autonomic optimization: a novel bio-inspired hybrid routing protocol for MANETs," *Future Generation Computer Systems*, vol. 81, pp. 505–513, 2018.
- [42] M. Zhang, Y. Zhou, W. Quan, J. Zhu, R. Zheng, and Q. Wu, "Online learning for IoT optimization: a Frank-Wolfe Adam-based algorithm," *IEEE Internet of Things Journal*, vol. 7, no. 9, pp. 8228–8237, 2020.
- [43] Y. Song, L. Liu, H. Ma, and A. V. Vasilakos, "A biology-based algorithm to minimal exposure problem of wireless sensor networks," *IEEE Transactions on Network and Service Management*, vol. 11, no. 3, pp. 417–430, 2014.
- [44] H. Wu, H. Ma, L. Xing, and G. Zheng, "Quality of video oriented and multi-meeting based routing algorithm for video data offloading," *IEEE Access*, vol. 6, pp. 36966–36976, 2018.
- [45] H. Ma, H. Wu, G. Zheng, and B. Ji, "Mcgr-pb: a multi-player cooperative game based routing for performance balancing in crowd sensing networks," *IEEE Access*, vol. 6, pp. 68440–68449, 2018.
- [46] T. Hao, J. Zhou, Y. Cheng, L. Huang, and H. Wu, "User identification in cyber-physical space: a case study on mobile query logs and trajectories," in *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 1–4, Burlingame, CA, USA, 2016.
- [47] R. Roedler, D. Kergl, and G. D. Rodosek, "Profile matching across online social networks based on geo-tags," *Advances in Nature and Biologically Inspired Computing*, vol. 419, pp. 417–428, 2016.
- [48] M. Qi, Z. Wang, Z. He, and Z. Shao, "User identification across asynchronous mobility trajectories," *Sensors*, vol. 19, no. 9, pp. 2102–2115, 2019.
- [49] D. M. Boyd and N. B. Ellison, "Social network sites: definition, history, and scholarship," *Journal of Computer-Mediated Communication*, vol. 13, no. 1, pp. 210–230, 2007.
- [50] S. Wang, X. Li, Y. Ye et al., "Anchor link prediction across attributed networks via network embedding," *Entropy*, vol. 21, no. 3, article e21030254, p. 254, 2019.
- [51] L. Xing, K. Deng, H. Wu, P. Xie, H. V. Zhao, and F. Gao, "A survey of across social networks user identification," *IEEE Access*, vol. 7, pp. 137472–137488, 2019.
- [52] G. Rossetti and R. Cazabet, "Community discovery in dynamic networks," *ACM Computing Surveys (CSUR)*, vol. 51, no. 2, pp. 1–37, 2018.
- [53] M. Yan, J. Sang, and C. Xu, "Unified YouTube video recommendation via cross-network collaboration," in *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pp. 19–26, New York, 2015.