symmetry

MDPI

Article

Behavioral Habits-Based User Identification Across Social Networks

Ling Xing [1,*], Kaikai Deng [1], Honghai Wu [1], Ping Xie [1] and Jianping Gao [2]

[1] School of Information Engineering, Henan University of Science and Technology, Luoyang 471023, China
[2] School of Vehicle & Transportation Engineering, Henan University of Science and Technology, Luoyang 471003, China
* Correspondence: xingling_my@163.com

Received: 10 August 2019; Accepted: 3 September 2019; Published: 5 September 2019

Abstract: Social networking is an interactive Internet of Things. The symmetry of the network can reflect the similar friendships of users on different social networks. A user's behavior habits are not easy to change, and users usually have the same or similar display names and published contents among multiple social networks. Therefore, the symmetry concept can be used to analyze the information generated by the user for user identification. User identification plays a key role in building better information about social network user profiles. As a consequence, it has very important practical significance in many network applications and has attracted a great deal of attention from researchers. However, existing works are primarily focused on rich network data and ignore the difficulty involved in data acquisition. Display names and user-published content are very easy to obtain compared to other types of user data across different social networks. Therefore, this paper proposes an across social networks user identification method based on user behavior habits (ANIUBH). We analyzed the user's personalized naming habits in terms of display names, then utilized different similarity calculation methods to measure the similarity of the features contained in the display names. The variant entropy value was adopted to assign weights to the features mentioned above. In addition, we also measured and analyzed the user's interest graph to further improve user identification performance. Finally, we combined one-to-one constraint with the Gale–Shapley algorithm to eliminate the one-to-many and many-to-many account-matching problems that often occur during the results-matching process. Experimental results demonstrated that our proposed method enables the possibility of user identification using only a small amount of online data.

Keywords: user identification; across social networks; display name; variant entropy value; interest graph

1. Introduction

In recent years, an increasing number of social networks have become indispensable tools for communication in our daily life. We can now share our ideas, status, location, etc. on social networks in real time and in ways not previously available to us. According to a 2019 statistical report [1], there are about 2.32 billion active users on Facebook and 1.098 billion active users on WeChat (https://weixin.qq.com/) every month. Moreover, since existing social network services cannot meet all of a user's needs simultaneously, each user tends to have multiple social network accounts [2]. The widespread application of social networks, however, also brings many problems, such as privacy leaks, hacker attacks, and network security. Because there is a certain connection between users on the social network, and the information generated by the user (profile data, behavior data, network structure) has certain symmetry, with this feature, we can identify the entity users behind multiple social network accounts.

As a result, information about a given user contained on any individual social network is incomplete. However, if we could identify the user's identity information across social networks, this would have important practical implications for many applications, including the ability to do the following:

1. Integrate user information from multiple accounts, enabling more accurate judgement of the user's hobbies, etc. and thus the ability to offer better recommendations and services [3];
2. Improve analysis and prediction of user behavior patterns in ways that cannot be achieved on a single social network [4];
3. Provide researchers with more complete user data [5];
4. Detect malicious users in a timely manner and provide targeted assistance to the network security field.

Existing works on user identification issues are discussed in the related works section below (Section 2). In brief, many of these existing works focus primarily on user profile information [6], such as username, gender, education, etc. Since major social networks first began to focus on user privacy protection, the difficulty of obtaining these types of profile information has increased. Furthermore, users' profile information may also be faked, which brings further challenges to user identification [7]. Some researchers use friend relationships to perform user identification [8]. As many users make their friend relationships public, the friend relationships between user accounts is very easy to access. However, the connections between friend relationships are generally sparse, with the result that user identification methods based on friend relationship networks also have certain limitations. Some researchers also utilize user-generated content to identify users [9,10]. The data published by users on social networks are easy to obtain and can facilitate a better analysis of user behavior data. User identification methods based on generated content can overcome the limitations of the above two methods.

In this paper, we mainly used display names and interest graphs for user identification purposes. A display name is the information filled in by users when they register social network accounts and takes the form of a string composed of characters, letters and numbers. This string contains rich redundant information and can reflect users' naming habits, meaning that we can extract this information for user identification purposes. Moreover, the content published by the user reflects the user's interests to a certain extent; thus, the user's interest graph can be well predicted by analyzing the content posted by the user. By combining the above two types of data, it is possible to achieve better identification performance.

The main contributions of this paper can be summarized as follows:

- We analyzed the redundant information contained in the display name and extract the length feature, character feature, and letter feature;
- We adopted the variant entropy value to assign weights to the features contained in the display name;
- We used the latent Dirichlet allocation (LDA) model to analyze the content posted by users and extract the user's interest graph;
- In order to improve user identification performance, we combined one-to-one constraint with the Gale–Shapley algorithm to optimize the user account matching results.

The remainder of this paper is organized as follows. Section 2 describes the current related works in this field. In Section 3, we define the problem of user identification. Section 4 introduces the proposed method of user identification, while Section 5 provides an experimental analysis of our proposed method. Finally, we summarize this paper in Section 6.

2. Related Works

Across social networks, user identification is of great importance in many domains, including personalized recommendation, information security, privacy protection, etc. Current works on this subject can be divided into three main categories: user profile data-based, network structure-based, and user-generated content-based methods.

2.1. User Profile Data-Based User Identification

User profile data refers to data that a user needs to enter or select when registering a social network account and include their username, gender, birthday, etc. In the process of filling out personal information, users are likely to refer to their previous registration information on other social networks. Studies show that users have similarities in profile data on different social networks, that is, there is symmetry in such data. Nowadays, many researchers are using this type of data to conduct research. Zafarani et al. [11] first proposed this method of user identification, which is generally utilized to add or remove prefixes and suffixes of appellations and to map usernames from one community to another for user identification purposes. Perito et al. [12] introduced a language-based model and a Markov chain technique by training the data of two social networks. Wang et al. [13] conducted in-depth research on username attributes and extracted thousands of features, including alphanumeric combination features, date features, etc. Li et al. [6] analyzed the differences in username choice across different social networks and constructed features that exploit information redundancies. The supervised machine learning method was adopted to further confirm the identified matching pairs.

Moreover, Vosecky et al. [14] proposed a method that transformed multiple attribute item information of users into n vectors, adopted different similarity calculation methods for each of a user's attribute items, and then selected different matching weights for different attribute items. Motoyama et al. [15] crawled and analyzed users' personal information on different social networks, represented it as a set of words, and then calculated the similarity between the words to obtain the similarity between different accounts. Raad et al. [16] designed a matching method based on the Friend-of-a-Friend (FOAF) vocabulary, transferred user profile data to this FOAF vocabulary, and then implemented a decision algorithm to obtain the similarity between two social accounts. Iofciu et al. [17] jointly considered usernames and user tags and utilized a simple subjective weighting method to weight them. Ye et al. [18] also proposed an objective weighting method based on subjective orientation to calculate the similarity among multiple user attributes. Li et al. [19] proposed an across social networks user identification model based on username and display name (UISN-UD), which contain rich information redundancy. The proposed method could conceivably reduce the use of attributes, as well as the degree of computational complexity. The most prominent advantage of this approach is that it both protects personal privacy and is highly accessible.

2.2. Network Structure-Based User Identification

User identification based on network topology information refers to methods in which the friend relationships between users are treated as equivalent to the network topology, allowing similarity matching between nodes to be performed. The network topology formed by users has certain symmetry on different social networks. Since the friend relationship formed by the user is basically fixed, identifying users on different social networks can be completed according to the number of shared identified friends. Narayanan et al. [20] were the first to prove that user identification could be achieved by relying on network topology information. Cui et al. [21] proposed combining users' profile information with the similarity of the graph to achieve mapping from an email network to a Facebook network; however, this mapping relationship was found to have a one-to-one mapping conflict problem. Kong et al. [22] transformed the problems identified in the literature [21] into prediction problems pertaining to directed links. Korula et al. [23] abstracted the user identification problem into a mathematical form, arguing that different social networks are generated by user graph

structure through probability. Tan et al. [24] proposed the concept of the hypergraph and designed a novel subspace learning method known as 'manifold alignment on hypergraph' (MAH). Zhou et al. [7] utilized the number of seed nodes shared by user nodes as a measure of similarity across different social networks, such that the ones with the largest similarity were selected for matching. Subsequently, Zhou et al. [25] designed an unsupervised scheme referred to as a friend relationship-based user identification algorithm without prior knowledge (FRUI-P). This algorithm extracts the friend features of each account in the social network as a feature vector, then calculates the similarities between all candidate users across the two social networks via in- and out-degrees. The main advantages of this method are that it does not need to know the seed nodes and can provide reliable prior knowledge for user identification.

2.3. User-Generated Content-Based User Identification

User identification based on user-generated data focuses primarily on the content published by users. Generally speaking, when users post social contents, they usually synchronize them with other social networks they hold. We can use the principle of symmetry to respectively analyze the data generated by users on different social networks, such as geographic locations, tags and status timestamps. Almishari et al. [26] took advantage of users' different writing styles to connect them across different online social networks, which verifies the linkability between different social networks. Nie et al. [5] subsequently proposed a dynamic core interest mapping (DCIM) algorithm that considers user topic model and topology structure based on user-generated content and ego-networks. Sha et al. [27] utilized statuses and comments posted by users to implement user identification across multiple social networks. Roedler et al. [28] used the timestamp information generated by users on social networks in conjunction with the location information generated by mobile devices to construct a personalized social behavior pattern in order to solve the user identification problem. Li et al. [29] designed a user-generated content-based user identification model (U-UIM), in which several algorithms are developed to measure the similarity of user-generated content (UGC) in terms of space, time and content dimensions. Moreover, supervised machine learning algorithms were also used to match users, which improved the comprehensive user identification performance.

In summary, in this paper, we use the redundant information contained in the user display name to analyze the relationship between accounts on different social networks, extract the length, character and letter features, and employ the variant entropy value to weight the features. Moreover, the interest graph can map users' behaviors and habits in a personalized way, which is also an important feature in the process of user identification. Finally, the Gale–Shapley algorithm was used to accurately match accounts on different social networks. The user data used in this work are highly accessible, protect user privacy, and enable the possibility of user identification using only a small amount of online data.

3. Problem Definition

On social networks, a display name is information that the user chooses to show to other users and is typically the attribute that best reflects the user's naming habits. A display name is usually an alphanumeric string, which is not necessarily unique. Moreover, when users post, forward, and like text content on social networks, this can intuitively reflect the user's interests; therefore, the user's interest graph is also an important factor for us to consider in the user identification process.

In this section, we introduce a method of across-social-network user identification based solely on display name and interest graph. Given two social networks S_1 and S_2, U_1^1 and G_1^2 denote a set of display names and interest graphs on social network S_1. and $u_1^{n1} \in U_1^1$, $u_1^{n2} \in G_1^2$, where u_1^{n1} and u_1^{n2} denote the display name and interest graph of the nth registered user in S_1.

Figure 1 presents an illustrated example to present this concept. User identification is the mining of physical users behind multiple different social network accounts. In other words, we need to determine whether (v_1^n, v_2^k) is the same user and whether (v_1^m, v_2^k) is a different user. We define E_{ij} as

the identified account pair. If $(v_1^n, v_2^k) \in E_{ij}$, this proves that the account pair (v_1^n, v_2^k) belongs to the same user in real life.

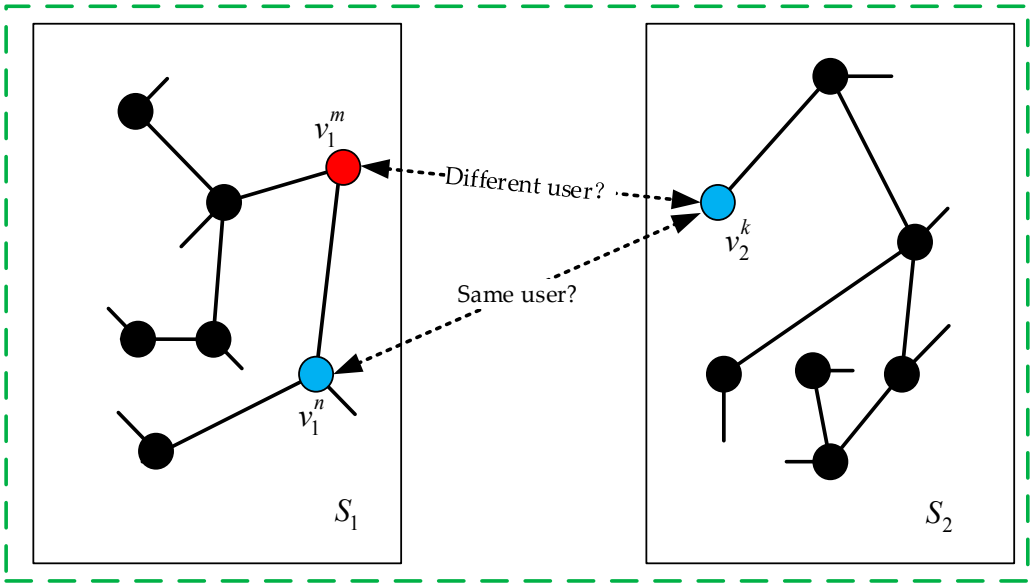


Figure 1. Example of across social networks user identification.

Without loss of generality, we assume that C and D are sets of display names and interest graphs respectively of two different user accounts across sites. The user identification solution attempts to discover all matching user pairs using an identification function $g(\cdot)$, such that:

$$g(\cdot) = \begin{cases} 1 & if\ C\ and\ D\ belong\ to\ the\ same\ person \\ 0 & otherwise \end{cases} \quad (1)$$

This identification function can be easily obtained using existing machine learning methods to train data. Therefore, our primary areas of focus are the feature analysis of display names and the construction of users' interest graphs. It is also important to note here that it is possible for a given display name to be shared by multiple users across different social networks, which has an impact on identification performance; in this paper, we ignored situations in which several people share the same display names.

4. The Method of User Identification

4.1. Display Name Analysis

When a user registers a new social network account, the display name selected by the user typically has substantial similarity to the display names selected by that user on other social networks. This phenomenon can highlight the user's behavior habits and reflect the user's identity information to a great extent. A display name is different from a username, which can be changed as often as the user likes. On some social networks (such as QQ (https://im.qq.com/), Foursquare, etc.), the username is a series of consecutive numbers. Therefore, the display names of users on multiple social networks may have symmetry due to behavioral habits, which is why this paper uses different similarity calculation methods for the features contained in the user display names.

This paper mainly measures and analyzes the length, character, and letter features of the display name in order to identify the user. When users register social network accounts, most users will utilize these three features to combine the display name. Li et al. [30] concluded that more than 45% of users have the same display name on different social networks, which provides an effective basis for the work of this paper. Accordingly, we used different similarity calculation methods to measure and analyzed the above three features.

4.1.1. Length Feature

There are several rules governing user display name choice. When the same user chooses a display name across different social networks, the length of the display names tends to be extremely similar. Assume that s_1 and s_1 are two display names, and that the length ratio is the ratio of the minimum and maximum values of the lengths of the two display names, which can be expressed by Equation (2):

$$R_{len} = \frac{\min(len(s_1), len(s_2))}{\max(len(s_1), len(s_2))} \tag{2}$$

Here, $len(s)$ represents the length of the display name. The length ratio is inversely proportional to the absolute value of the length difference, and $R_{len} \in (0, 1]$. When the length ratio is 1, it indicates that the two display names have the same length.

4.1.2. Character Feature

A Display name is composed of strings in different social networks. Therefore, we can combine the character features of a string to calculate the similarity between display names. In this subsection, the longest common substring is mainly used to obtain the similarity between two strings. The chosen measurement method is defined as the ratio of the length of the longest common substring to the minimum string length. The similarity is then proportional to the ratio between the two display names. Supposing that we have two substrings, l_1 and l_2, the calculation formula is as in Equation (3):

$$R_{lcs} = \frac{len(lcs(l_1, l_2))}{\min(len(l_1, l_2))}, 0 \le R_{lcs} \le 1 \tag{3}$$

where $lcs(l_1, l_2)$ denotes the longest common substring of the strings l_1 and l_2, while $len(l_1, l_2)$ denotes the length of strings l_1 and l_2.

4.1.3. Letter Feature

Letters are also a feature often used when users choose a display name. The same display name has a consistent letter distribution; for two similar display names, their letter distribution is also similar. For example, the display name "movie star" and the display name "star movie" have the same letter distribution. Since the number of possible letters is very large overall, we only consider the 26 English letters here. We measureD letter distribution similarity using cosine similarity.

Cosine similarity is mainly used to measure the similarity between two vectors. By calculating the frequency at which each English letter appears in the display name, we obtain a vector of display names, the calculation formula of which is as in Equation (4):

$$\cos\theta = \frac{\sum_{i=1}^{26} A_i \times B_i}{\sqrt{\sum_{i=1}^{26} A_i^2} \times \sqrt{\sum_{i=1}^{26} B_i^2}} \tag{4}$$

Here, A_i and B_i denote the frequency at which the ith letter appears in each display name. Since we only used 26 English letters, the range of i is [1,26]. The larger the $\cos\theta$ value, the greater the similarity between the two display names.

To better explain the vector formed by the display name, we shall use the display name "jacka" as an example: The corresponding display name vector would be [2,0,1,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0], that is, the vector is defined in terms of the frequency of letters appear.

4.1.4. Weight Assignment Based on Variant Entropy Value

We can obtain the similarity of each feature by analyzing the features extracted from the above display name. In the process of user identification, the extent to which each feature contributes to the identification result may be different; for example, the length feature is less important than the other two features when considering a user's display name choice. Therefore, we need to assign weights to the extracted features to improve user identification performance.

The traditional expert subjective weighting method is tightly coupled with the attribute domain, and the resulting algorithm is less robust. Moreover, the objective weighting method relies on a large amount of sample data, which is poor in terms of versatility and participation. Accordingly, in order to better solve the above problems, we adopt a weight assignment method based on variant entropy values to weight the extracted features. In information theory, an entropy value can reflect the order of information and the amount of information contained. Therefore, information entropy can be used to evaluate the importance of each feature to user identification. According to the definition of information entropy, when a system is in different states, the probability of occurrence of each state k is $p_{ik}(k = 1, 2, \ldots, m)$, with m denoting the output states of the source. In this paper, we used the concept of information entropy to represent m as the number of features contained in the display name. The information entropy calculation formula can thus be expressed as:

$$E_i = -\sum_{k=1}^{m} p_{ik} \times \log p_{ik} \tag{5}$$

where E_i denotes the information entropy of the kth state.

In Equation (5), p_{ik} is defined as the probability of occurrence of feature similarity. Its calculation formula is as follows:

$$p_{ik} = \frac{v_i^{jk}}{\sum\limits_{k=1}^{m} v_i^{jk}} \tag{6}$$

where v_i^{jk} represents the similarity of the ith feature between the account k in the target network and the account j in the source network. According to the above analysis, Formula (5) can be rewritten as:

$$E_i^{jk} = -\sum_{k=1}^{m} (v_i^{jk} / \sum_{k=1}^{m} v_i^{jk}) \times \log (v_i^{jk} / \sum_{k=1}^{m} v_i^{jk}) \tag{7}$$

Since the entropy value is inversely proportional to the weight, the variant entropy value Q_i^{jk} can be constructed as follows:

$$Q_i^{jk} = \frac{1}{E_i^{jk}} \tag{8}$$

Through the above derivation, we can obtain the weight assignment of each feature as follows:

$$w_i^{jk} = Q_i^{jk} / \sum_{i=1}^{m} Q_i^{jk} \tag{9}$$

The specific process for assigning weights to each feature in the display name is shown in Algorithm 1.

Finally, we can derive the weight-based similarity between the two display names as follows:

$$sim_{display} = \sum_{i=1}^{m} (w_i^{jk} \times v_i^{jk}) \tag{10}$$

Algorithm 1: Display Name Feature Weight Assignment

Input: Source network account feature vector F_C, feature vectors $\{F_k\}_{k=1}^{m}$ for all accounts in the target network, feature vector F_D to be matched account in the target network
Output: w_i^{jk} (the weight of the ith feature of accounts j and k)
1: For each F_k in $\{F_k\}_{k=1}^{m}$(m represents the number of all accounts to be matched in the target network)
2: for $i = 1$ to n
3: Calculate display name similarity v^{jk} of accounts C and D by using equations (2) (3) (4)
4: end
5: for $i = 1$ to n
6: The attribute weights of display name features are assigned using Equation (5) (6) (7) (8)(9)
7: end
8: Return w_i^{jk}

4.2. User-Published Content Analysis

When users register for social network accounts, their behavior data are posted on the corresponding social sites. At the same time, users will also comment, repost, and like/'thumbs up' the content published by other users. Many of the users' interests will change to some extent over time; however, some long-term interests of users are far less subject to change. This provides a new path for us to identify the user. The user's interest will not change because of different social networks. Instead, they will focus on the same topics of interest in different social networks. Therefore, the interest graphs generated by users on different social networks have symmetry and similarity to a certain extent, which can well achieve across social networks user identification. Related terms are defined as follows:

Definition 1 (Interest graph): *The content posted by the user in social networks reflects the long-term interests of the users, which can be defined as the interest graph.*

Definition 2 (Interest factor): *By analyzing the user's interest, it can be seen that a certain interest of the user changes continuously over a period of time, and this interest can be defined as an interest factor.*

Definition 3 (Node set): *A virtual account registered by a user on a different social network is equivalently a node in the process of user identification. All users on social networks form a node set.*

Definition 4 (Edge set): *The connection relationship between nodes constitutes edge sets, which indicates the degree of relationship between the nodes.*

The user's interest graph tends to be stable for a long time. If a change occurs, the user's data on other social networks will change accordingly. The principle behind this form of user identification involves mining the user's interest graph; we used the LDA model to obtain the user's topics of interest.

The basic idea is that each document (user published content) can be considered equivalent to a mixed distribution of a series of topics, so that a three-layer Bayesian model of "document-topic-word" can be constructed. Each document in the document set is categorized via probability distribution. According to the document generation rules and explicit data in the LDA model, the topic distribution is derived via expected value propagation. The method of generating a document via the LDA model is represented in Figure 2.

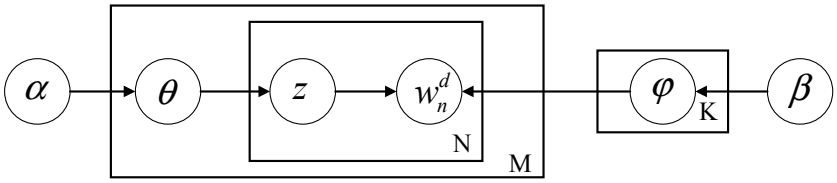


Figure 2. Latent Dirichlet allocation (LDA) model architecture.

Here, φ is the word distribution, θ is the topic probability distribution of document, α is the parameter of the Dirichlet distribution of θ, β is the parameter of the Dirichlet distribution of φ, z denotes the topic of the word, and w_n^d denotes the nth word in the dth document. Moreover, M denotes the number of documents, N denotes the length of the documents, and K denotes the number of topics.

We define $\theta_{i,t} = (\theta_{1,i,t}, \ldots, \theta_{K,i,t})^T$ as the probability distribution of the articles published by the ith person at time t on k topics. $\varphi_k = \left(\varphi_{1,k}, \ldots, \varphi_{v,k}\right)^T$ is the probability distribution of the subject k in the dictionary space formed by v words, v represents the subject words generated by all topics, while α and β are hyper-parameters.

For the article published by the ith person at time t, the generation process is as follows:

1. The abovementioned introduction to the basic knowledge and related symbols of LDA. For user-generated documents, the prior distribution of the topic is a Dirichlet distribution. In other words, for any document d, the probability distribution $\theta_{i,t} \sim Dir(\alpha)$ of the document on the k topics is generated from the Dirichlet distribution;
2. For the cth word in the document:
   (a) We first need to show the distribution of topics corresponding to the cth word, then the specific subject of its expression is derived from a multivariate distribution: $z_{i,t,c} \sim multi(\theta_{i,t})$;
   (b) Next, we should find out the specific words that correspond to the topic, generate a concrete word that expresses the subject from a multivariate distribution: $w_{i,t,c} \sim multi(\theta_{z_{i,t,c}})$;
3. Generate a probability distribution $\varphi_k \sim Dir(\beta)$ of the topic k on all words from the Dirichlet distribution.

Based on the above document generation process, the joint likelihood function of all documents can be obtained. The Gibbs sampling method can then be used to solve the model in order to obtain the estimated values of $\theta_{i,t}$ and φ_k, and the formulae are as follows:

$$\theta_{i,t} = \frac{n_d^{(k)} + \alpha_k}{\sum_{k=1}^{K} n_d^{(k)} + \alpha_k} \tag{11}$$

$$\varphi_k = \frac{n_k^{(c)} + \beta_c}{\sum_{c=1}^{v} n_k^{(c)} + \beta_c} \tag{12}$$

where $n_k^{(c)}$ denotes the number of times the word c appears in the topic k, and $n_d^{(k)}$ denotes that the document d corresponds to the count of the topic k.

Through the above steps, the probability distribution of the document on the K topics and the probability distribution of the topic k on the v words can be solved, thereby obtaining the user's topic vectors.

Moreover, in acknowledgement of the fact that the user's behavioral habits may change constantly over time, we defined a time interval window Δt to estimate the dynamic changes in user interest. The user's topics of interest can be represented by a matrix B, as shown below:

$$B = \begin{bmatrix} p_1(\theta) & \Delta t_1 \\ p_2(\theta) & \Delta t_2 \\ \cdots & \cdots \\ p_m(\theta) & \Delta t_m \end{bmatrix} \tag{13}$$

Here, $p(\theta)$ denotes the user's topic distribution, and Δt denotes the interval time.

After obtaining the user's topic distribution, it is necessary to distinguish the user's interest graph and interest factor. We used Kullback-Leibler (KL) divergence to calculate the similarity of the distribution of topics at different time periods. KL divergence is an asymmetry calculation method for calculating the degree of difference in probability distributions. For the vectors $p(i)$ and $q(i)$ of two probability distributions, KL divergence is calculated as follows:

$$D(P\|Q) = \sum_i p(i) \cdot \log(\frac{p(i)}{q(i)}) \tag{14}$$

We defined the similarity of the user's interest distribution at different Δt as $Sim(P\|Q)$, which is calculated by Equation (15):

$$Sim(P\|Q) = [D(P\|Q) + \lambda]^{-1} \tag{15}$$

Here, λ is a minimum value, which is mainly used to avoid the denominator being 0.

By calculating the similarity between the topics, the user's interest graph can be determined through comparison with the set threshold T. If the similarity is greater than the threshold, the topic is determined to be an interest graph. Then, a new matrix B_1 can be reconstructed using the user's interest graph:

$$B_1 = \begin{bmatrix} p_{11}(\theta) & \Delta t_{11} \\ p_{22}(\theta) & \Delta t_{22} \\ \cdots & \cdots \\ p_{mn}(\theta) & \Delta t_{mn} \end{bmatrix} \tag{16}$$

After obtaining the user's interest graph matrix, we can use cosine similarity to calculate the similarity of the interest graphs between different accounts. The specific calculation formula is as follows:

$$Sim_{li}(B_i, B_j) = \frac{\sum_{t=\Delta t_{11}}^{\Delta t_{mn}} B_i \times B_j}{\sqrt{\sum_{t=\Delta t_{11}}^{\Delta t_{mn}} B_i^2} \times \sqrt{\sum_{t=\Delta t_{11}}^{\Delta t_{mn}} B_j^2}} \tag{17}$$

where B_i and B_j are interest graph vectors formed by user accounts on two different social networks.

The process is outlined in more detail in Algorithm 2 below.

Algorithm 2: Similarity Calculation of User Interest Graph

Input: Behavior data of user accounts i and j, related parameter settings including α, β, Δt and T.
Output: Interest graph similarity between accounts i and j.
1: Set the time interval window Δt
2: The topic distribution $p(\theta)$ of user account is calculated via the LDA model
3: Form the topic matrix B
4: Calculate the KL divergence of the user topic by Equation (14)
5: Calculate the similarity between topics by Equation (15)
6: The user's interest graphs are obtained by comparing the threshold T
7: Reconstitute topic matrix B_1
8: Interest graph similarity of accounts i and j is calculated using cosine similarity
9: Return $Sim_{ij}(B_i, B_j)$

4.3. User Account Matching

As shown in Figure 3, given two different social networks S_1 and S_2, and three user accounts v_1^n, v_1^m, and v_2^k where $v_1^n, v_1^m \in S_1$, $v_2^k \in S_2$, the display names and interest graphs of v_1^n, v_1^m, and v_2^k are (u_1^{n1}, u_1^{n2}), (u_1^{m1}, u_1^{m2}) and (u_2^{k1}, u_2^{k2}). Let us select accounts v_1^n and v_2^k as examples. This pair of accounts is mapped to node r_{nk}; therefore, matching user accounts v_1^n and v_2^k transforms into a classification problem. If y_{nk} is the classification result of r_{nk}, then when $y_{nk} = 1$, accounts v_1^n and v_2^k can be assumed to belong to the same user; otherwise, the two accounts belong to different individuals. Therefore, we can address this problem via the supervised machine learning method. For the identified accounts, $\forall n, k\left(v_1^n, v_2^k\right) \in E_{ij}$, $y_{nk} = 1$. We can obtain the feature vector X_{nk} from r_{nk}, which denotes the information contained in the display names and interest graphs. We can then construct training data (X_{nk}, y_{nk}), which can be used to train a supervised classifier.

If there are no constraints on the user matching results, cases in which $y_{nk} = 1$ and $y_{mk} = 1$ will occur, leading to a one-to-many and many-to-many problem with the identification results. To avoid this problem, we used Equations (16) and (17) to achieve a one-to-one constraint:

$$z(y_{mn}, y_{kl}) = \begin{cases} 1 & m = k \; or \; n = l \\ 0 & otherwise \end{cases} \tag{18}$$

$$\forall v_1^m \in S_1, \sum_{v_{2 \in G_2}} y_{mk} = 1 \tag{19}$$

If $z(y_{mn}, y_{kl}) = 1$, this means that only one pair of user accounts has a value of 1; that is, the final user identification results should be satisfied (Equation (19)). We are influenced by the concept of stable marriage matching in handling the problem of user identification. Thus, the Gale–Shapley algorithm [31] is adopted to enable the above problems to be better solved. The specific process is outlined in Algorithm 3 below.

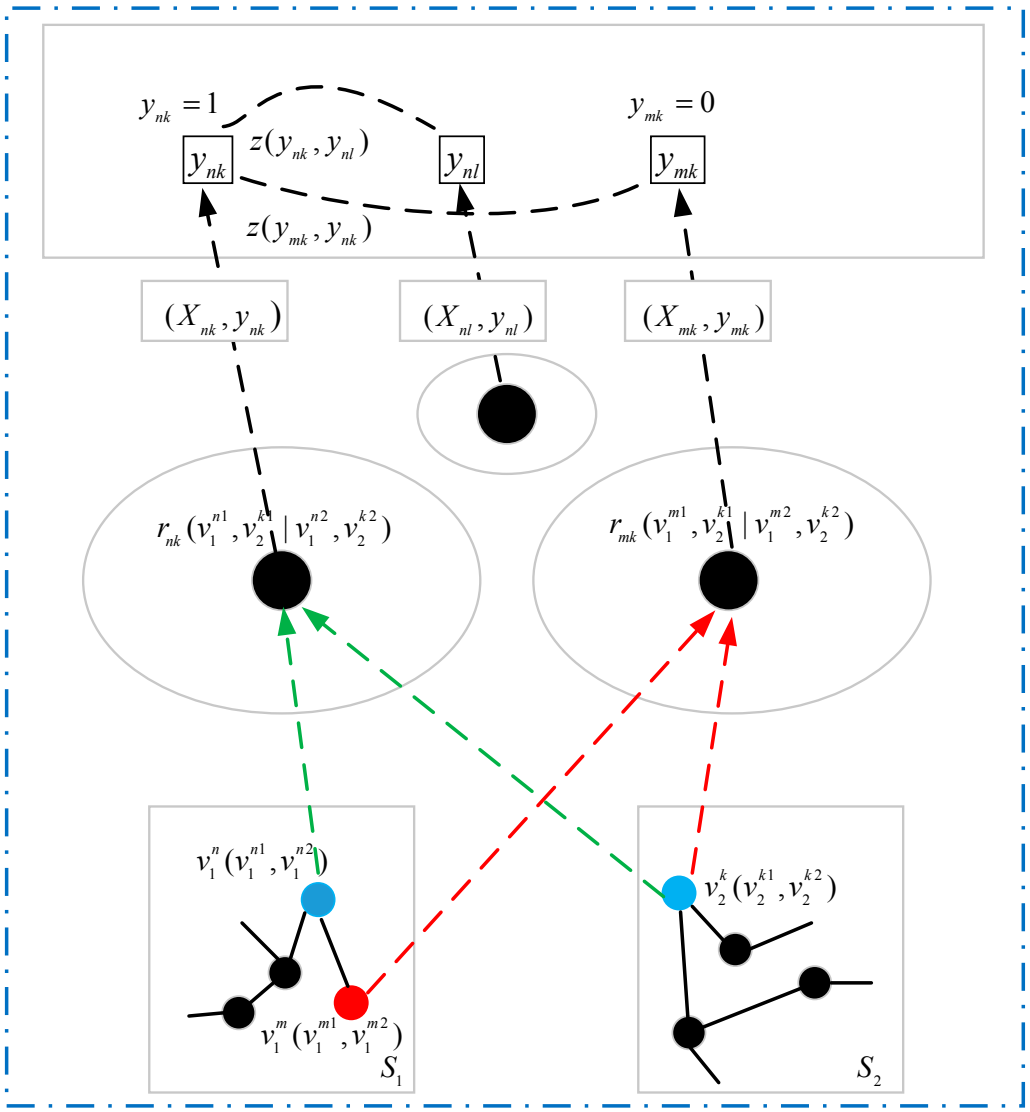


Figure 3. User account identification framework.

Algorithm 3: User Matching with One-To-One Constraint

Input: $\{s(X_{nk})\mid 0\le n\le|S_1|,0\le k\le|S_2|\},R=\phi$
Output: $R=\{(v_1^n,v_2^k)\mid y_{nk}=1\wedge equation\ (19)\}$
1: For each user account v_1^n, v_2^k belonging to S_1 and S_2
2: Two probability sets $\{rank^1\}$ and $\{rank^2\}$ are formed by the classifier, respectively
3: while $\exists v_1^n\in S_1$ or $\{rank^1\}\neq\phi$
4: Select v_2^k from $\{rank^2\}$, $v_2^k=\text{argmax}(probability)$
5: If v_2^k not matched
6: Add (v_1^n,v_2^k) to R
7: Else
8: Compare the priorities of user account v_1^m and v_1^n in $\{rank^2\}$(Assume that v_2^k and v_1^n are matched accounts.)
9: If $v_1^m>v_1^n$
10: Remove (v_1^n,v_2^k) from R
11: Add (v_1^m,v_2^k) to R
12: Else
13: ignore (v_1^m,v_2^k)
14: Return R

5. Experimental Results and Analysis

Experiments were conducted to better illustrate the effectiveness of the method proposed in this paper. All experiments were performed on a computer with 8G memory and a 2.4GHz CPU.

5.1. Dataset analysis

To obtain the users' display names and published contents, we first need to know the social networks that users have registered. There are many ways to obtain user information through social networks, including questionnaire survey and web crawler. We used the programming software Python to crawl the data needed for the experiment from two different social networks: 'Sina Weibo' (https://weibo.com/) and 'Toutiao' (or 'Today's Headlines') (https://www.toutiao.com/). Sina Weibo is similar to Twitter in that it is a popular and public platform on which users can share their blog posts anytime and anywhere. Today's Headlines is a news client that can provide accurate, personalized recommendation services for users based on their interests, age, and other information. To illustrate the effectiveness of the proposed method, the specific crawl data are shown in Table 1.

Table 1. User information collection statistics.

Datasets	Data Type	Number
Sina weibo	Display name User published content	2000 Six months
Douban	Display name User published content	2000 Six months

We crawled the user's display name and published content. In the process of processing user display names, we found that the display names of users on different social networks have certain similarities, which indicates the importance of analyzing user behavior habits in this paper. When acquiring user content, some Douban users display their Weibo identity or URL in their profile, which can be used as ground truth of user identification between two social networks. Since the user usually synchronizes to other social networks when posting content, the performance of the user identification is further improved by analyzing the user's interest graph. We crawled and analyzed the data of users on these two social networks for six months. We assigned 75% of the data to the training set and the remaining 25% to the test set. Moreover, since some users of these social networks have blank display names, we ignored these users in the experiment.

5.2. Evaluation Metrics

When analyzing the effectiveness of different user identification methods and comparing their advantages and disadvantages, the most commonly used evaluation metrics are precision rate, recall rate, and F-measure (F1). The definitions of these metrics can be expressed as Equations (20)–(22):

$$precision = \frac{tp}{tp + fp} \tag{20}$$

$$recall = \frac{tp}{tp + fn} \tag{21}$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \tag{22}$$

Here, tp denotes account pairs that belong to the same user and are correctly matched. fp denotes the number of pairs where the two corresponding accounts belong to different users but are identified as a matching pair (false positives), while fn denotes the number of users that are not matched but are in fact the same users (false negatives).

5.3. Selection of Experimental Parameters

Many parameters are used to optimize the user identification performance in the process of obtaining the user interest graph: These include threshold T, KL divergence, time interval window Δt, etc. We explain the setting of these parameters in detail below.

5.3.1. Threshold Setting of Interest Graph

The user's interest graph can be obtained using the crawled user data. Over the course of the experiment, the LDA model was used to analyze the user's topic distribution. We set the values of α, β, and K as $\alpha = 50/K$, $\beta = 0.01$ and $K = 20$. The users' interest graphs are defined via a reasonable threshold T. If the threshold T is too large, then topics that should be part of the interest graph are mistakenly identified as interest factors; conversely, if the threshold T is too small, the topics of the interest graph will be mistaken for interest factors. The setting range of the threshold T in the experiment is [0.1, 1]. Figure 4 presents the impact of different thresholds on identification performance. We can clearly see that when the threshold is less than 0.7, both F1 and precision rate constantly increase; when the threshold is greater than 0.7, moreover, the F1 and recall rate tend to decrease gradually, while the precision rate tends to be stable. Therefore, the impact of threshold changes on evaluation metrics is fully considered, and the final threshold value is set to 0.7.

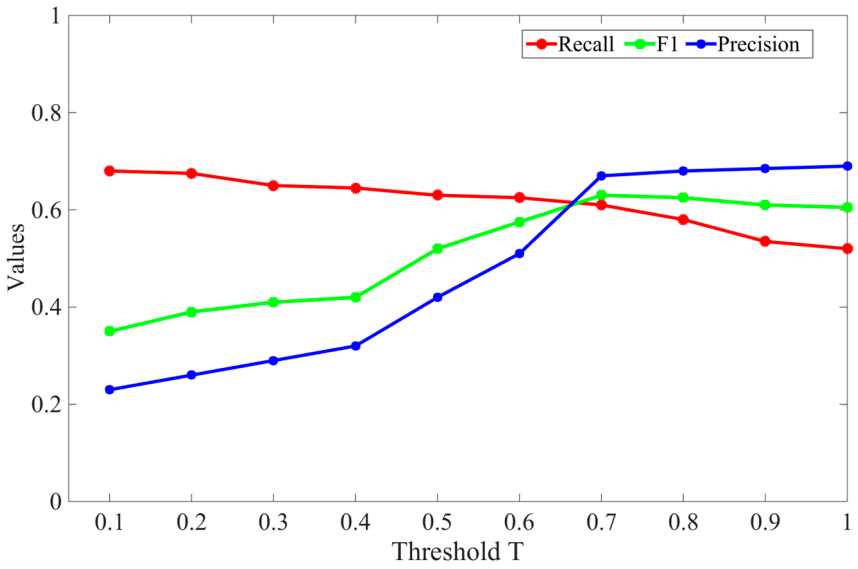


Figure 4. Threshold selection.

5.3.2. KL Divergence Distribution of User Interests

The KL divergence reflects the relational degree between the various topics of interest to a user: the smaller the value, the higher the relational degree between the two topics. We adopted KL divergence to judge the relationship between different topics in the process of building the interest graph. Based on the above threshold setting, we can use KL divergence to calculate the difference between users' interest graphs and interest factors. From Figure 5, we can see that the users' interest graphs are very stable compared with an interest factor over a period of time. According to the changes in user interest shown in Figure 4, we can clearly see that the interest graph can be used for user identification.

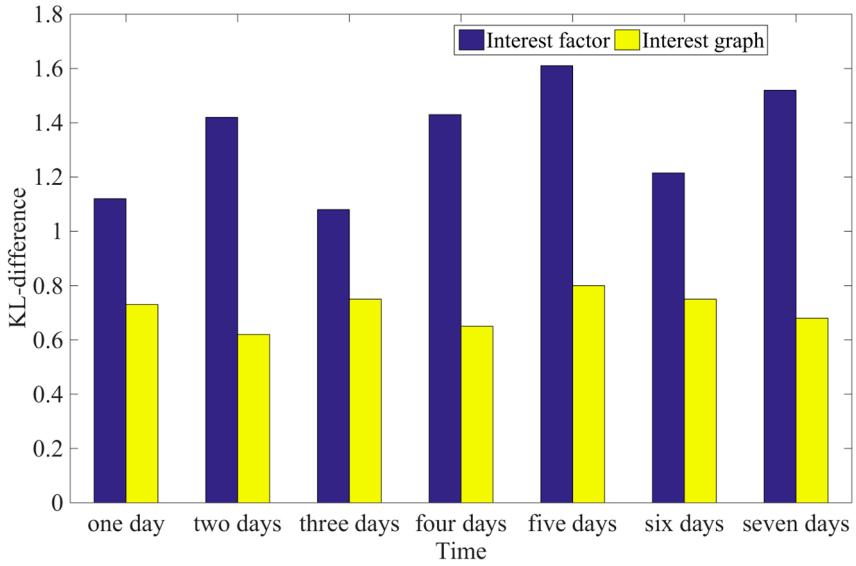


Figure 5. KL difference of user interests.

5.3.3. Time Interval Window Selection

Another important parameter requiring our attention is the time interval window. The appropriate choice of this parameter is a prerequisite for obtaining a user interest graph. The classification of user interests can be more effectively implemented by defining the time interval window Δt at a reasonable size: If the value of Δt is too small, it is difficult to obtain a user's interest graph, while if the value of Δt is too large, the amount of redundant user information will increase, which will cause certain difficulties when attempting to draw the distinction between interest graph and interest factors. As shown in Figure 6, when the value of Δt is three days, the value of F1 reaches the maximum; therefore, $\Delta t = 3$ is finally selected as the set value.

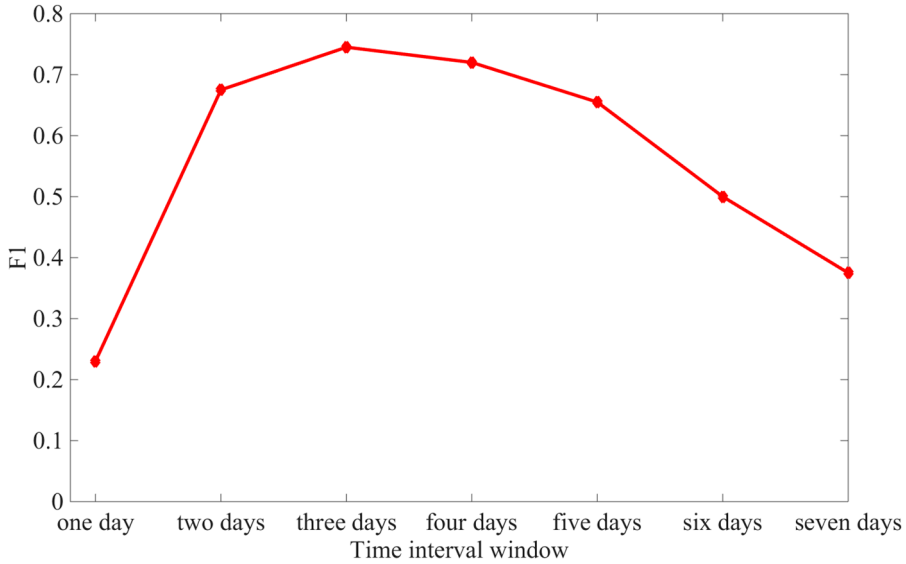


Figure 6. The setting of time interval window Δt.

5.4. Impact of Data Used on Identification Performance

In order to effectively analyze and compare the influence of display names and user interests on the identification results, we evaluated five schemes to verify the effectiveness of the proposed method. The five schemes are as follows: $ANIUBH_{nodn}$ (the display name is not used during the

identification process), $ANIUBH_{noig}$ (the interest graph is not used in the identification process), $ANIUBH_{now}$ (the display name is not assigned a weight during the identification process), *DPUI* (the information entropy-based weight is assigned to the display name during the identification process) [32], and *ANIUBH* (the proposed method). In the interests of clarity, these schemes are hereafter represented by the letters A, B, C, D, and E, respectively. Each method is tested on the crawled data set.

5.4.1. Impact of User Data on User Identification

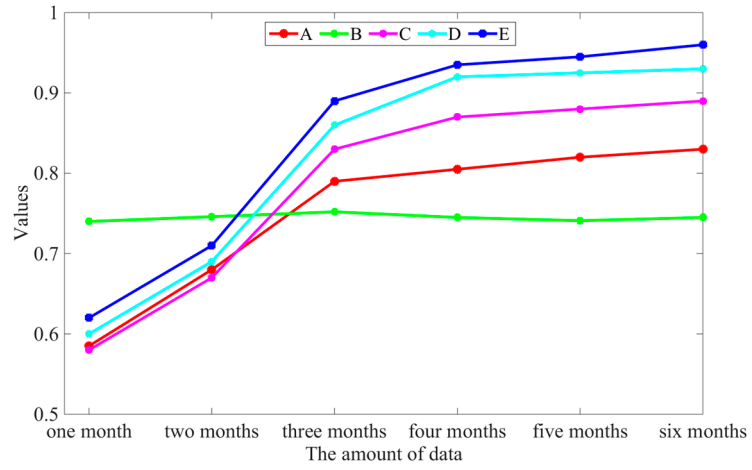
In this subsection, we mainly compare the impact of user data on user identification results. We used two main types of data for user identification: display name and user-published content. The identification performance of the A, B, and C schemes (see above) is primarily compared and analyzed here. As shown in Figure 7, identification performance based solely on the display name is better than that of the other two schemes. The reason for this is that a user's interest graph is difficult to create when the amount of available user data is small. With an increasing amount of user data, the user identification evaluation metrics also increase. However, we can clearly see that the curve of the B scheme has hardly changed. This phenomenon indicates that the user's display name is time-independent over time.

5.4.2. Impact of Weight on User Identification

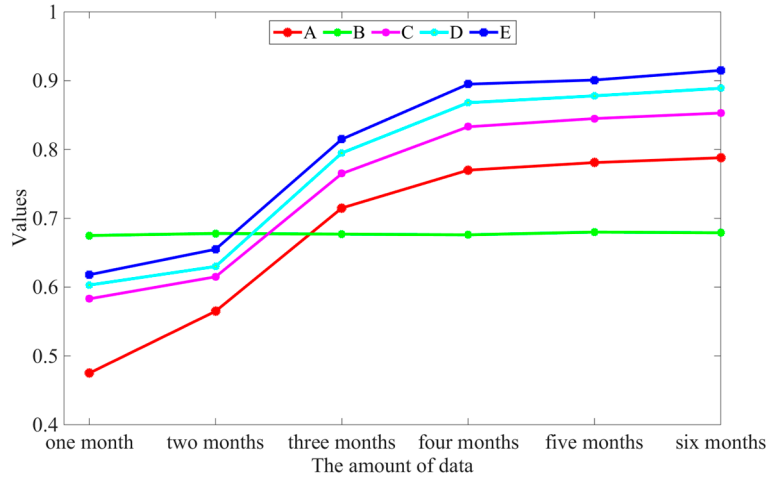
In Section 4, we presented a detailed analysis of the importance of weighting to user identification. In this subsection, we mainly analyze the identification performance comparison of the schemes C, D, and E. As shown in Figure 7, we can clearly see the difference in identification performance between the three schemes. The D scheme adopts an information entropy-based method for weight assignment; however, it can be seen from the experimental results that the method of variant entropy proposed in this paper is better than the weight assignment method of the D scheme. As the amount of training data continuously increased, the evaluation metrics of E improved greatly compared with the other four methods. Moreover, from the figure, we can see that when the user data extends over than four months, the evaluation metrics tend to be stable. This phenomenon demonstrates that the user data trained for four months can basically achieve a good identification performance. Taking this into account can allow a reduction in the amount of calculations required in the training data process to some extent.

5.5. Complexity Analysis

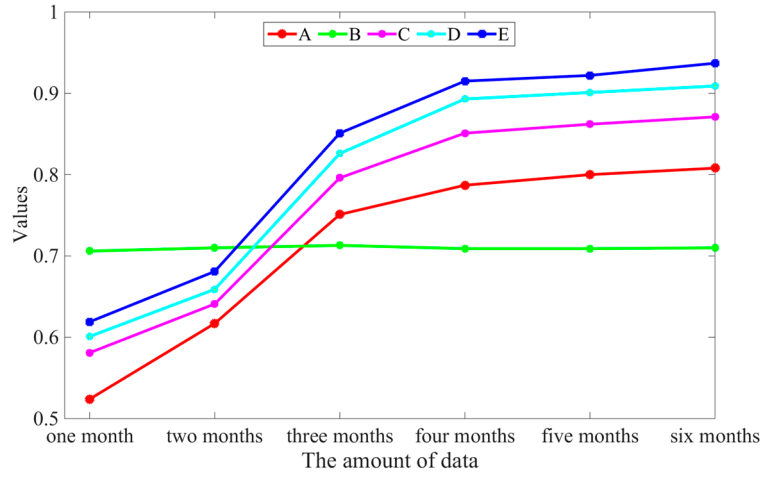
Generally speaking, the best user account match is to satisfy the one-to-one constraint. The complexity of Algorithm 3 is $o(n)$ when Equation (17) is satisfied, where $n = \min(|S_1|, |S_2|)$. The worst situation is that in which an account on social network S_1 matches all user accounts on social network S_2. In this situation, the complexity of Algorithm 3 is $o(n^2)$. Therefore, we can conclude that the complexity of Algorithm 3 is between $o(n)$ and $o(n^2)$. In the interests of comparison, the complexity of user identification based on the network structure proposed in [25] is $o(n^2)$. The effectiveness of the proposed method can be demonstrated by analyzing the complexity of the above algorithm.



(**a**)



(**b**)



(**c**)

Figure 7. Comparison of evaluation metrics of five schemes. (**a**) Precision comparison. (**b**) Recall comparison. (**c**) F1 comparison.

6. Conclusions

In the era of big data, user data can be crawled through many channels. The display name and published content of users on social networks generate rich redundant information. Compared to other forms of user data, these two types of data are far less limited by privacy protections, have higher accessibility, and generate data with symmetry and similarity; however, the most important point is that they can also reflect the user's behavior habits. We first extracted the length, character, and letter features from the user's display name and assigned weights to the extracted features. We considered that a user's interest graph can map onto the user's real-life habits. We used the LDA model to analyze the user's topic distribution and thereby obtained the user's interest graph. The data analyzed above were then fused and combined one-to-one constraint with the Gale–Shapley algorithm to optimize the user account matching results. From the experimental analysis section, it can be concluded that the user data and related algorithms employed in the present paper are of great help to identification performance.

Author Contributions: Conceptualization, L.X. and K.D.; methodology, L.X. and K.D.; validation, L.X. and K.D.; data curation, L.X.; writing—original draft preparation, K.D.; writing—review and editing, L.X., K.D., H.W., P.X., and J.G.; supervision, L.X., H.W., P.X., and J.G.; project administration, L.X., H.W., P.X.

Funding: This research is supported by National Natural Science Foundation of China (Grant No. 61771185, Grant No. 61772175, Grant No. 61801171), Science and Technology Research Project of Henan Province (Grant No. 182102210044, Grant No. 182102210285), Key Scientific Research Program of Henan Higher Education (Grant No. 18A510009), and Postdoctoral Science Foundation of China under Grant 2018M632772.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Most Famous Social Network Sites Worldwide as of April 2019, Ranked by Number of Active Users. Available online: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ (accessed on 10 August 2019).
2. Liu, J.; Zhang, F.; Song, X.; Song, Y.I.; Lin, C.Y.; Hon, H.W. 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, Rome, Italy, 4–8 February 2013; pp. 495–504.
3. Zheng, J.X.; Li, D.Y.; Kumar, S.A. Group user profile modeling based on neural word embeddings in social networks. *Symmetry* **2018**, *10*, 435. [CrossRef]
4. Li, C.; Lin, S. Matching users and items across domains to improve the recommendation quality. In Proceedings of the KDD, New York, NY, USA, 24–27 August 2014; ACM: New York, NY, USA, 2014; pp. 801–810.
5. Nie, Y.; Jia, Y.; Li, S.; Zhu, X.; Li, A.; Zhou, B. Identifying users across social networks based on dynamic core interests. *Neurocomputing* **2016**, *210*, 107–115. [CrossRef]
6. Li, Y.J.; Peng, Y.; Ji, W.L.; Zhang, Z.; Xu, Q.Q. User identification based on the display name across online social network sites. *IEEE Access* **2017**, *5*, 17342–17353. [CrossRef]
7. Zhou, X.P.; Liang, X.; Zhang, H.Y.; Ma, Y.F. Cross-platform identification of anonymous identical users in multiple social media networks. *IEEE Trans. Knowl. Data Eng.* **2016**, *28*, 411–424. [CrossRef]
8. Shu, K.; Wang, S.; Tang, J.; Zafarani, R.; Liu, H. User identity linkage across online social networks: A review. *ACM SIGKDD Explor.* **2017**, *18*, 5–17. [CrossRef]
9. Deng, K.K.; Xing, L.; Zheng, L.S.; Wu, H.H.; Xie, P.; Gao, F.F. A user identification algorithm based on user behavior analysis in social networks. *IEEE Access* **2019**, *9*, 47114–47123. [CrossRef]
10. Deng, K.K.; Xing, L.; Zhang, M.C.; Wu, H.H.; Xie, P. A multiuser identification algorithm based on internet of things. *Wirel. Commun. Mob. Comput.* **2019**, *2019*, 6974809. [CrossRef]
11. Zafarani, R.; Liu, H. Connecting corresponding identities across communities. In Proceedings of the International Conference on Weblogs and Social Media, San Jose, CA, USA, 17–20 May 2009; Volume 9, pp. 354–357.
12. Perito, D.; Castelluccia, C.; Kaafar, M.A.; Manils, P. How unique and traceable are usernames. In *International Symposium on Privacy Enhancing Technologies Symposium*; Springer: Berlin/Heidelberg, Germany, 2011; Volume 6794, pp. 1–17.

13. Wang, Y.B.; Liu, T.W.; Tan, Q.F.; Shi, J.Q.; Guo, L. Identifying users across different sites using usernames. *Procedia Comput. Sci.* **2016**, *80*, 376–385. [CrossRef]
14. Vosecky, J.; Hong, D.; Shen, V.Y. User identification across multiple social networks. In Proceedings of the 2009 First International Conference on Networked Digital Technologies, Ostrava, Czech, 28–31 July 2009; pp. 360–365.
15. Motoyama, M.; Varghese, G. I seek you: Searching and matching individuals in social networks. In Proceedings of the 11th International Workshop on Web Information and Data Management, Hong Kong, China, 2 November 2009; pp. 67–75.
16. Raad, E.; Chbeir, R.; Dipanda, A. User profile matching in social networks. In Proceedings of the 13th International Conference on Network-Based Information Systems, Takayama, Japan, 14–16 September 2010; pp. 297–304.
17. Iofciu, T.; Fankhauser, P.; Abel, F.; Bischoff, K. Identifying users across social tagging systems. In Proceedings of the 5th International AAAI Conference on Weblogs and Social Media, Barcelona, Spain, 1 January 2011; pp. 522–525.
18. Ye, N.; Zhao, L.; Dong, L.; Bian, G.; Liu, E.; Clapworthy, G.J. User identification based on multiple attribute decision making in social networks. *China Commun.* **2013**, *10*, 37–49.
19. Li, Y.J.; Peng, Y.; Zhang, Z.; Yin, H.Z.; Xu, Q.Q. Matching user accounts across social networks based on username and display name. *World Wide Web.* **2018**, *22*, 1075–1097. [CrossRef]
20. Narayanan, A.; Shmatikov, V. De-anonymizing social networks. In Proceedings of the 30th IEEE Symposium on Security and Privacy, Los Alamitos, CA, USA, 17–20 May 2009; Volume 1, pp. 173–187.
21. Cui, Y.; Pei, J.; Tang, G.T.; Luk, W.S.; Jiang, D.X.; Hua, M. Finding email correspondents in online social networks. *World Wide Web.* **2013**, *16*, 195–218. [CrossRef]
22. Kong, X.; Zhang, J.; Yu, P.S. Inferring anchor links across multiple heterogeneous social networks. In Proceedings of the 22nd ACM International Conference on Information and Knowledge Management, San Francisco, CA, USA, 27 October–1 November 2013; pp. 179–188.
23. Korula, N.; Lattanzi, S. An efficient reconciliation algorithm for social networks. *Proc. VLDB Endow.* **2014**, *7*, 377–388. [CrossRef]
24. Tan, S.L.; Guan, Z.Y.; Cai, D.; Qin, X.Z.; Bu, J.J.; Chen, C. Mapping users across networks by manifold alignment on hypergraph. In Proceedings of the 28th AAAI Conference on Artificial Intelligence, Québec, QC, Canada, 27–31 July 2014; pp. 159–165.
25. Zhou, X.P.; Liang, X.; Du, X.Y.; Zhao, J.C. Structure based user identification across social networks. *IEEE Trans. Knowl. Data Eng.* **2018**, *30*, 1178–1191. [CrossRef]
26. Almishari, M.; Tsudik, G. Exploring linkability of user reviews. *Comput. Secur.-ESORICS* **2012**, *7459*, 307–324.
27. Sha, Y.; Liang, Q.; Zheng, K.J. Matching user accounts across social networks based on users message. *Procedia Comput. Sci.* **2016**, *80*, 2423–2427. [CrossRef]
28. Roedler, R.; Kergl, D.; Rodosek, G.D. Profile matching across online social networks based on geo-tag. *Adv. Nat. Biol. Inspired Comput.* **2016**, *419*, 417–428.
29. Li, Y.J.; Zhang, Z.; Peng, Y.; Yin, H.Z.; Xu, Q.Q. Matching user accounts based on user generated content across social networks. *Future Gener. Comput. Syst.* **2018**, *83*, 104–115. [CrossRef]
30. Li, Y.J.; Peng, Y.; Zhang, Z.; Wu, M.J.; Xu, Q.Q. A deep dive into user display names across social networks. *Inf. Sci.* **2018**, *447*, 186–204. [CrossRef]
31. Dubins, L.; Freedman, D. Machiavelli and the Gale-Shapley algorithm. *Am. Math. Mon.* **1981**, *88*, 485–494. [CrossRef]
32. He, Z.M.; Li, W.J. Research on user identification across multiple social networks based on preference. In Proceedings of the 2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems, Nanjing, China, 23–25 November 2018; pp. 122–128.