IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received March 1, 2019, accepted March 29, 2019, date of current version April 18, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2909089

A User Identification Algorithm Based on User Behavior Analysis in Social Networks

KAIKAI DENG[1], LING XING[1], LONGSHUI ZHENG[2], HONGHAI WU[1], PING XIE[1], AND FEIFEI GAO[3], (Senior Member, IEEE)

[1] School of Information Engineering, Henan University of Science and Technology, Luoyang 471023, China
[2] School of Information Engineering, Southwest University of Science and Technology, Mianyang 621000, China
[3] Institute Artificial Intelligence, Tsinghua University, Beijing 100084, China

Corresponding author: Ling Xing (xingling_my@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61771185, Grant 61772175, and Grant 61801171, in part by the Science and Technology Research Project of Henan Province under Grant 182102210044 and Grant 182102210285, in part by the Key Scientific Research Program of Henan Higher Education under Grant 18A510009, and in part by the Postdoctoral Science Foundation of China under Grant 2018M632772.

ABSTRACT The precision of the conventional user identification algorithm is not satisfactory because it ignores the role of user-generated data in identity matching. In this paper, we propose a frequent pattern mining-based cross-social network user identification algorithm that analyzes user-generated data in a personalized manner. We adopt the posterior probability-based information entropy weight allocation method that improves the precision rate and recall rate compared to the empirical weight allocation method. The extensive simulations are provided to demonstrate that the proposed algorithm can enhance the precision rate, recall rate, as well as the F-Measure (F1).

INDEX TERMS User identification, frequent pattern, cross-social network, information entropy.

I. INTRODUCTION

In the past decade, many types of social networks have emerged and have contributed immensely to the large volume of real-world data on social behaviors. Due to the diversity of social networks, people tend to use different social networks for different purposes [1]. For instance, RenRen, a Facebook-style social network in China, is used to share blogs, while wechat is used to share personal statuses. However, there is no direct link between individual social network accounts and a complete cross-social network map is difficult to obtain. Cross-social network user identification technology has been proposed to identify of the physical users of each social network [2].

User identification, also called user matching, has been achieved to via examining similarities of user profile information attributes, including username, birthday, gender, location, etc. [3]–[9]. These attributes could be easily copied and easily forged for different purposes (including malicious users). Some other researchers have leveraged similarities of location, posting time, and writing style to identify users [10]–[15]. However, these techniques are plagued with limitations since location data is difficult to obtain and writing style is difficult to extract from short sentences. Some other methods applied network topology for user identification, which mainly relies on the user's circle of friends [11], [16], [17]. Due to the heterogeneity of social networks in practical applications, which will affect the similarity calculation between nodes.

Social network data can be divided into user profile data, user-generated data, and user-associated data. User profile data refer to the data that the user needs to enter or select when they register for their social network account, including their username, gender, age, etc; User-generated data refer to the user's blog posts and their publication time; User-associated data refer to the data that are associated with a user on a social network and other users, such as likes, comments, and repostings [18]. In this paper, the first two types of data are mainly referred to as user data. User behavior analysis is based on user-generated data that largely reflect the characteristics of individual users. Some user data occupy a more important for user identification than others, such as web links, user blogs, etc. For example, the users on the two social networks have the same username, then it is likely that they are the same user. Conversely, some other user data do not have an important status, such as gender, age, etc.

The associate editor coordinating the review of this manuscript and approving it for publication was Pasquale De Meo.

For example, only the users of two social networks have the same gender, it is difficult to say that they are the same user.

In this paper, we propose a frequent pattern mining-based cross-social network user identification algorithm (FPM-CSNUIA) to analyze user-generated data. We adopt the *SimFunc()* function to calculate the similarities between all user data by selecting the vector of the account in different social networks. Since each user data has a different effect on user identification, we propose a posterior probability-based information entropy weight allocation method to assign weights for user data, while the weight-based similarity vector is constructed for user identification.

II. RELATED WORKS

Cross-social network user identification technology is important in many research fields, such as information retrieval, artificial intelligence, cyber security, etc. Current studies on social network user identification can be divided into three categories: user profile information-based, network structure-based, and user-generated data-based.

A. USER PROFILE INFORMATION-BASED USER IDENTIFICATION

User profile information-based user identification mainly utilize attribute information, such as username, gender, and birthday. This method mainly determines whether the profile information of two accounts in different social networks match or not. For example, Raad *et al.* [19] proposed a friend of a friend (FOAF) attribute matching method, which transfers the user's configuration information to the FOAF vocabulary and forms a unified FOAF format to obtain the similarity between the attributes. Cortis *et al.* [20] designed an identity recognition algorithm that assigns weights to individual attributes in user profile information and calculates the similarity among attributes in terms of both the grammatical and semantic aspects. Iofciu *et al.* [21] introduced an approach that measures the distance between user profiles. Able *et al.* [22] aggregated user profile information and matched users across social networks. Li *et al.* [23] analyzed the differences in naming patterns on different social networks, constructed features that exploit information redundancies, and adopted supervised machine learning to identify users. Esfandyari *et al.* [24] proposed a overlapping attribute items method to select user information that makes the trained model more applicable, but the method increases the computational complexity. Li *et al.* [25] proposed a UISN-UD model that uses username or display names to contain rich information redundancy to match users, which could reduce the use of attributes in the user identification process as well as the computational complexity degree. The most prominent advantage is that it does not involve personal privacy and is highly accessible. The comprehensive evaluation index exceeds 90%, [25] provides a new idea for user identification. [19]–[25] purely relying on user profile information-based schemes have limitations, because some attributes are easily replicated in large-scale social networks.

B. NETWORK STRUCTURE-BASED USER IDENTIFICATION

Network structure-based studies on user identification mainly utilize the user's circle of friends to identify user. Zhou *et al.* [1] utilized the number of seed nodes shared by user nodes as a measure of similarity across social networks where they selected the ones with the largest similarity for matching. Kong *et al.* [11] calculated the similarity between user nodes using the extended Adamic-Adar index and Jaccard correlation coefficient. Bartunov *et al.* [16] constructed the objective function with combining attribute information and network structure information, which optimizes the function to obtain the matching pair. Zhou *et al.* [17] designed an unsupervised scheme, termed friend relationship-based user identification algorithm without prior knowledge (FRUI-P), which extracts the friend feature of each user, and then calculates the similarities of all the candidate users between the two social networks. The advantage of [17] is that it is not necessary to know the seed nodes and reliable prior knowledge for identifying users. Since the heterogeneity between nodes in the actual social network and the influence of heterogeneity is ignored in the calculation process, this type of approach has an impact on precision of user identification.

C. USER GENERATED DATA-BASED USER IDENTIFICATION

User-generated data-based user identification mainly relies on the timing of blog postings, the usage habits of special symbols, and word usage. This method mainly determines whether the generated information of two accounts in different social networks match or not. Almishari and Tsudik [10] took advantage of the different writing styles of users to connect them to different online social networks. Kong *et al.* [11] proposed multi-network anchoring (MNA) to map users; comprehensively calculate the similarity between user's social, spatial, time, and textual information; and examine the problem with stable user matching. Goga *et al.* [12] used the geographic locations, timestamps, and writing habits of user's published statuses for user identification. Han *et al.* [13] proposed that each geographic coordinate point should be represented as a corresponding semantic position. The user's trajectory can be represented by the text composed of the semantic position, with the LDA model being used to represent the user's topic distribution. Then, the similarity of the user trajectories is calculated to determine whether the two users are the same. Li *et al.* [14] designed a UGC-based user identification model (U-UIM), where several algorithms are developed to measure the similarity of UGC in space, time and content dimensions. Moreover, supervised machine learning algorithms are used to match users, which improved the comprehensive performance of user identification. Since user-generated data is personalized, the analysis of user behavior data for identifying users across social networks is a good choice.

D. DATA PREPROCESSING

Since user data is stored as strings, the similarity value of a corresponding user data can be obtained by calculating the similarity between string sequences. A common similarity calculation is as follows:

1) DICE COEFFICIENT

When calculating the multi-valued strings n_i and n_j, the sum of the two times of the intersection information divided by the sum of the elements of n_i and n_j is the two string Dice coefficients, which are computed as follows:

$$SinFunc_{Dice}(n_i, n_j) = 2\frac{|n_i \cap n_j|}{|n_i| + |n_j|}. \tag{1}$$

For example, in two multi-valued attribute strings *"dynamic music movie"* and *"music reading experience"*, the intersection information is *"music"*. Hence, the similarity is $2 \div 6 \approx 0.33$.

2) COSINE SIMILARITY

Cosine similarity can be used when the similarity between two vectors is calculated [26]. Assuming that $X = [X_1, X_2, \ldots, X_n]$ and $Y = [Y_1, Y_2, \ldots, Y_n]$ are two n-dimensional vectors. The cosine of the Angle between X and Y is calculated as:

$$\cos(X, Y) = \frac{\sum_{i=1}^{n}(X_i \times Y_i)}{\sqrt{\sum_{i=1}^{n}(X_i^2)} \times \sqrt{\sum_{i=1}^{n}(Y_i^2)}} \tag{2}$$

If the angle between the two vectors is closer to "1", then the similarity between the two users is higher. If the angle between the two vectors is closer to "0", then the similarity between the two users is smaller.

III. USER IDENTIFICATION METHODS ACROSS SOCIAL NETWORKS

A. OVERALL FRAMEWORK OF THE METHOD

Cross-social network user identification is to find out the physical users behind multiple virtual social network accounts, and is important in many research fields. This paper mainly analyzes user profile data and user behavior data, and combines the above two data types to further improve the precision of user identification. The attributes with missing data in user profile information are filled with other user data, and is carried on the generalization processing. The similarity value is obtained by the related user data calculation method, and is compared with the threshold to obtain a similarity vector composed of "0" or "1". For user behavior data, we mainly analyzes the user's blogs, state timestamps and special symbols, and then constructs a weight-based similarity vector together with the user profile data. The weight-based similarity of the accounts between the two social networks is analyzed to determine whether the users have the same identity. The overall framework of the propose method is shown in Fig. 1.

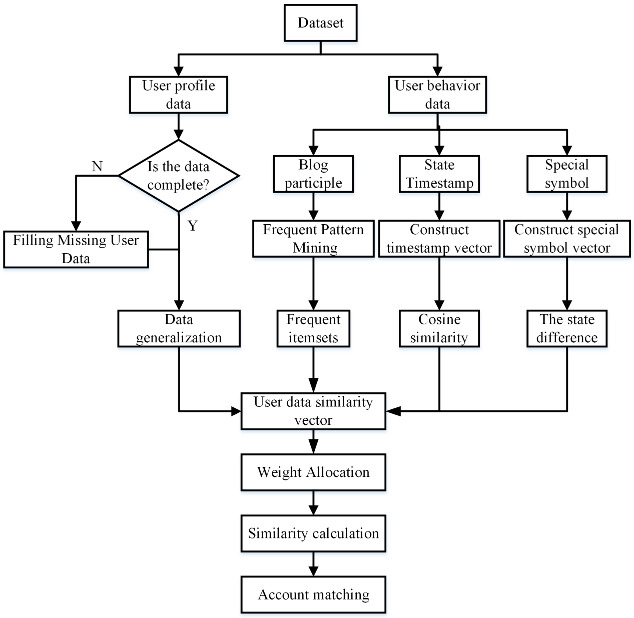


FIGURE 1. Overall block diagram of the system.

B. USER PROFILE DATA ANALYSIS

For one thing, due to privacy protection or laziness, users may have missing data when filling in user profile information. For another, there are differences in the format and integrity of user profile data on social networks. Therefore, there are several techniques for analyzing user profile data.

1) FILLING MISSING USER DATA

Due to the incompleteness of the information in the user profile data, an appropriate filling method should be adopted for missing user data. Note that not every dimension of user profile data is suitable for processing, because some profile information can not be populated, such as web links, mottos, emails, etc.

a: SIMILARITY FILLING

We select the n users with the highest relational degree between other users and users with missing data and take the mode number to fill the user profile data. The relational degree can be obtained from (3), where c_{AB} denotes mutually commenting, r_{AB} denotes reposting, and then l_{AB} denotes thumbing. Such three types of user behavioral information are sequentially assigned the weights "3", "2", and "1" in this paper. The formula is as follows:

$$R_{AB} = 3\sum c_{AB} + 2\sum r_{AB} + \sum l_{AB} \tag{3}$$

b: SPECULATED FILLING

The missing data can be inferred from other attributes. For example, the gender of the user can be inferred from the blog posted by the user. The Bayesian formula is written as follows.

$$p(m|w_i) = \frac{p(w_i|m)p(m)}{p(w_i)} \tag{4}$$

where $p(m|w_i)$ denotes the probability that the user is identified as male when the word occurs, $p(w_i|m)$ denotes the probability of word occurrence in all males, $p(m)$ denotes the probability of the user being male, and $p(w_i)$denotes the probability of word occurrence.

In order to reduce the complexity of calculating $p(w_i|m)$, we calculate the conditional independence naïve hypothesis. Therefore, the Bayesian formula is rewritten as:

$$p(m|w_i) = \frac{p(w_1|m)p(w_2|m)\cdots p(w_n|m)p(m)}{p(w_1)p(w_2)\cdots p(w_n)} \tag{5}$$

The statistical results of the training set can be used to derive the probability of (5). The prediction results of the corresponding attributes can be obtained via the Bayesian formula.

2) USER PROFILE DATA SIMILARITY CALCULATION

Since the formats of the user profile data entered when users register their social network accounts are inconsistent, they need to be generalized before calculating the similarities between user attributes. The user data being processed is more suitable for similarity calculation. The Dice coefficient and Cosine similarity are commonly used to calculate the similarity between user attributes. Each dimension is provided with a threshold value when calculating similarity. If the similarity of each dimension is greater than the corresponding threshold value, then it will return "1"; otherwise, it will return "0". If "1" is returned, then user accounts can be treated as the same user on this dimension. After all dimensions are calculated, a vector composed of "0" or "1" will be constructed, and will be used to build the weight-based similarity vector.

C. USER BEHAVIOR DATA ANALYSIS

User behavior analysis leverages user-generated data to process similarities between users. User-generated data refer to the textual data and the timestamp accompanying it. That is, the time at which the user's blog posts and blog updates are generated. Since the user behavior habit is not easy to change and hide [27], we propose a frequent pattern-based method to analyze user text behavior habits, leverage Cosine similarity to measure the similarity between the user's special symbols, and measures the similarity between the timestamp of the user's personal statuses through the state difference.

1) FREQUENT PATTERN MINING ALGORITHMS

Due to the different living environments and personalities of users, significant differences exist in their writing styles and the special symbols. Since it is difficult to subjectively find some special rules, the data mining algorithms are required to discover these hidden association rules [28]. This paper uses frequent pattern mining to analyze user-generated data with the following steps:

Step 1: Word segmentation is performed on the blog posts published by the user on social networks. After the word segmentation, blog post consists of many words and constructs a set of transaction T_1. Therefore, all blog posts constitute transaction set $D = \{T_1, T_2, \cdots, T_k\}$.

Step 2: Transaction set D is scanned to count the number of occurrences of each word, and then generate a candidate set C_1. We set that the minimum support count of the one-item set is two, and delete frequent items with a support degree of less than 2 to obtain the one-item set L_1.

Step 3: The candidate set C_2 can be obtained by connecting the one-item set L_1. When the item set is greater than one, the minimum support is one. The transaction set D is scanned to filter out the frequent items that do not meet the necessary support for obtaining L_2.

Step 4: Similarly, Connect to L_{k-1} execute pruning strategy to generate candidate item set C_k, then the transaction set D is scanned to count each item in C_k. Sets of *three*-item, *four*-item, $\cdots$, k-item are sequentially generated until L_k fails to meet the minimum support. The specific process is summarized in Algorithm 1.

Algorithm 1: Frequent Pattern Mining Algorithm

Input: Transaction set D, minimum support threshold min _ sup.
Output: Frequent item set L_k.

```
$L_1$=Frequent one-itemsets
for ($k = 2; L_k = \phi; k++$)
  $C_k = Apriori\_gen\{L_k\}$
    for each transaction set $T \in D$ do
      $C_T = subset(C_k, T)$
        for each candidate set $c \in C_T$ do
          $c = c + 1$
        end for
    end for
end for
$L_k$= candidates in $C_k$ with support $\geq$ min _ sup
return $L_k$
```

2) SPARK-BASED FREQUENT PATTERN MINING IMPLEMENTATION PROCESS

The frequent pattern mining algorithm is implemented on the Spark distributed platform. The main process can be divided into two steps.

Step 1: Import word segmentation of all user's blogs as the original dataset. Perform a flatMap on the original dataset to form RDD0 and then execute MapToPair on the original transaction set to get the frequent one item set RDD1. The data format of RDD1 is <item, 1>. Then execute parallel reduceByKey on RDD1 to get RDD2<item, count>, where value is the support count of this item on the transaction. Finally, a filter operation is performed on RDD2, and the item set that does not satisfy the minimum support count is filtered to obtain RDD3<item, count> that is a frequent one item set. The specific process is shown in FIG. 2.

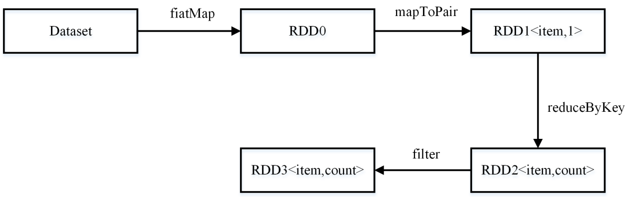


FIGURE 2. The first step of Apriori algorithm.

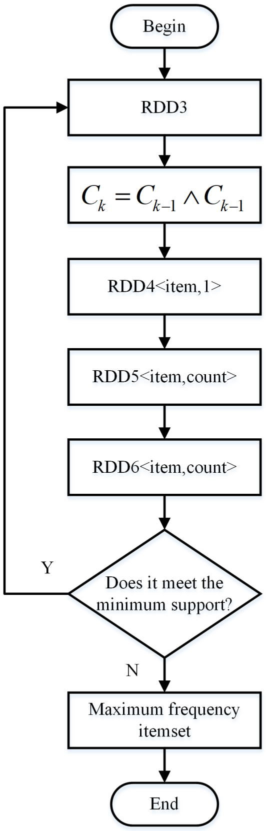

Begin
RDD3
$C_k = C_{k-1} \wedge C_{k-1}$
RDD4<item,1>
RDD5<item,count>
RDD6<item,count>
Y
Does it meet the minimum support?
N
Maximum frequency itemset
End

FIGURE 3. The second step of Apriori algorithm.

Step 2: This step is a process of iteratively performing connection and pruning operations. As shown in FIG. 3, the obtained frequent one item set is input into the driver to obtain the set C_1. Then the connection operation is performed on C_1 to obtain the candidate two item set C_2, and the obtained C_2 is distributed to each slave node for traversal. If this candidate set is included, a key-value pair <itemset,1> is output, resulting in RDD4<itemset,1>. Run the reduceByKey operation on RDD4 to generate RDD5<itemset, count>, perform a filter operation on RDD5, filter out the item set that does not satisfy the minimum support to form RDD6<itemset, count>, and then use RDD6 to generate three items. The iterative calculation is performed in turn until a new item set cannot be acquired, and the algorithm execution ends.

After data mining, the similarity of user blog data can be calculated. Effective frequent item sets requir a lot of statistical data, but some users have very few blog data. In this paper, the one-item set is used as a calculation indicator, the similarity can be calculated even if users have less blog data.

The similarity between the user-generated data of user U_A of social network M and user U_B of social network N is calculated as follows:

$$Sim_{U_A U_B} = \sum_{E_i \in U_A \cap U_B} (1 + C_{U_A}(E_i)) \times (1 + C_{U_B}(E_i))^{C(E_i)} \quad (6)$$

where $C_{U_A}(E_i)$ denotes the support degree count of the frequent item set E_i of user U_A, $C_{U_B}(E_i)$ denotes the support degree count of frequent item set E_i of user U_B, $C(E_i)$ denotes the item set number of E_i, and "1" is added to avoid a high-frequency item set. According to the historical data results. If $Sim_{U_A U_B} > 5000$, then return "1"; otherwise, return "0".

3) STATE TIMESTAMP SIMILARITY CALCULATION

The dynamic time generated by users on the social network shows the personalized characteristics of users, and has a positive effect on user identification. The user's time per day can be divided into 24 time intervals. The average dynamic number of each time interval of the user can be obtained to construct a user timestamp vector of 24 dimensions. The similarity calculation formula of state timestamp is expressed as:

$$Sim_t(p, q) = \sum_{i=1}^{n} |u_{pi} - u_{qi}| \quad (7)$$

where u_{pi}, u_{qi} denotes the average timestamp of users p and q, n denotes 24 dimensions of the users timestamp vector.

According to the statistical results, If Sim_t is less than 0.1, then the user accounts belong to the same user; otherwise the user accounts do not belong to the same user.

4) SPECIAL SYMBOLS SIMILARITY CALCULATION

The use of specific symbols also reflects the user's personalized characteristics, which is an important indicator to measure user identity information. Other features may cause the user to have certain differences in the use of special symbols, such as the user's personality, working environment, gender, etc. We quantize the word features of each user into special symbol vectors base on the frequency with which the user utilizes special symbols in the blog post, and then leverage Cosine similarity (2) to calculate the similarity of user pairs from different social networks. According to the previous training set, the $cos_{x,y}$ should be generally greater than 0.98 to determine that user accounts belong to the same user.

D. THE POSTERIOR PROBABILITY-BASED INFORMATION ENTROPY WEIGHT ALLOCATION ALGORITHM

1) WEIGHT ANALYSIS OF USER DATA

The similarity vector can be obtained by calculating the similarities between user profile data and user-generated data. Since user data have a different influence on the degree of

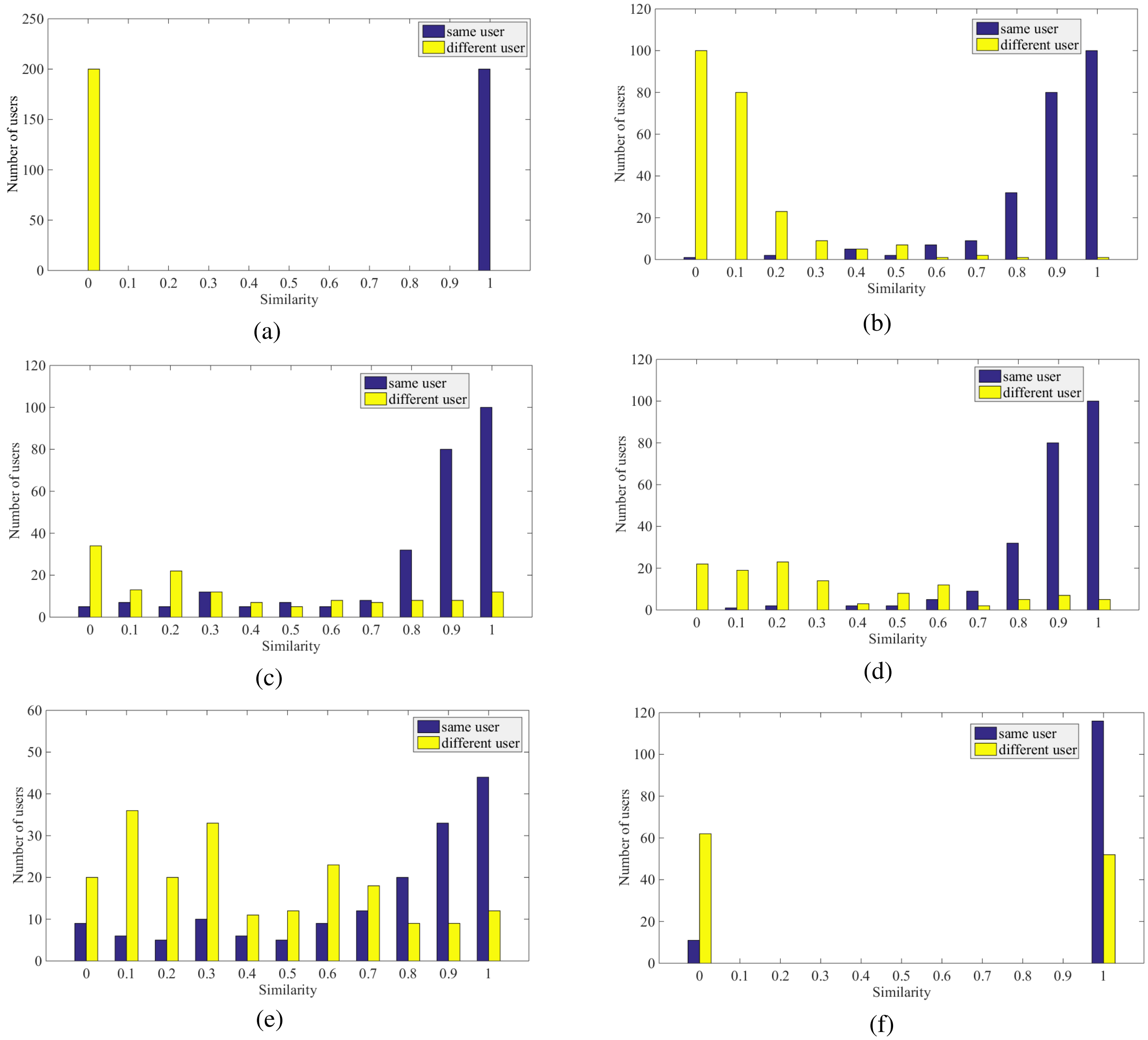


FIGURE 4. User Data Similarity Distribution. (a) URL map. (b) Username distribution map. (c) State timestamp distribution map. (d) Frequent pattern distribution map. (e) Interests distribution map. (f) Gender map.

user identification, it is necessary to calculate the weight of user data. Fig. 4 shows the performance of a single data in user identification. The URL, username, status timestamp, and frequent pattern have different distributions of similarity between the same user and different user when performing user matching. These data are highly distinguishable. Therefore, the weight allocation should be relatively large. When users match in terms of their interests and gender, the similarity distribution between the same user and different user is small. Hence, the weight distribution should be small. Since each user data has different effects on user identification, it is necessary to assign corresponding weights to each user data.

2) WEIGHT ALLOCATION ALGORITHM

The traditional weight allocation algorithm has some limitations in terms of robustness and universality. Therefore, we propose the posterior probability-based information entropy weight allocation algorithm that takes the weight allocation into consideration from various perspectives and improves precision in identifying users.

In information theory, the entropy value reflects the degree of information disorder. The smaller its value is, the more orderly the information is, and the more information it contains; on the contrary, the more disordered the information is, the less information it contains. Therefore, information entropy can be used to evaluate the effectiveness of user data. According to the definition of information entropy, for any random variable, the formula is:

$$E(x) = -\sum_{x \in X} p(x) \log(p(x)) \tag{8}$$

where *p(x)* is the value probability for the user data.

In order to assign more reasonable weight to user data. The posterior probability of user data is further calculated

TABLE 1. User data type classification.

User data	Data type
username	user profile data
gender	user profile data
birthday	user profile data
hometown	user profile data
biography	user profile data
URL	user profile data
interest	user profile data
current residence	user profile data
language	user profile data
religious belief	user profile data
political views	user profile data
motto	user profile data
marital status	user profile data
website	user profile data
email	user profile data
education	user profile data
professional experience	user profile data
frequent pattern	user behavior data
state timestamp	user behavior data
specific symbols	user behavior data

on the basis of information entropy, which helps improve the precision of user identification. By combining the posterior probability and information entropy, the weight of the user data is expressed as:

$$W(x) = -p(y_s|s)\sum_{x\in X} p(x)\log(p(x)) \tag{9}$$

where $p(y_s|s)$ is the posterior probability of the user data. As shown in Table 1, the user data information contains 20 types, of which 17 types are user profile data and 3 types are user behavior data. In the process of filling the user profile data, some user data cannot be filled and are represented by null. The similarity of each user data is set with a relevant threshold. If the similarity exceeds the threshold, then output "1"; otherwise output "0".

The ith user data information of account j is defined as d_i^j. The user similarity vector is defined as $V(F_A, F_B) = (v_1^{AB}, v_2^{AB}, \cdots, v_n^{AB})$, where v_i^{AB} represents the similarity between the ith user data of user F_A and user F_B. The main purpose of user identification is to identify the same user in different social networks by calculating the similarity in user data between accounts. In order to improve the precision of user identification, we propose a posterior probability-based information entropy weight allocation method to assign reasonable weights for user data. Finally, a weight-based similarity calculation formula is constructed as follows.

$$Sim_w(F_A, F_B) = \sum_{i=1}^{n}(w_i^{AB} \times v_i^{AB}) \tag{10}$$

In general, the threshold is set as 0.5. If Sim_w is greater than 0.5, then the user accounts belong to the same user; otherwise the user accounts do not belong to the same user.

E. USER ACCOUNT MATCHING

Through the analysis of user profile data and user-generated data, the similarity of user data is respectively calculated, and the weight corresponding to each user data is assigned to form weight-based similarity vector. User matching can be divided into three steps:

Step 1: Select account: If all the user data of the accounts F_A and F_B are calculated, then the computational complexity will be high. Therefore, it is necessary to filter the target account in another network according to the condition C: filter accounts by username.

Step 2: Obtain candidate set: The candidate set can be obtained after filtering accounts, and the similarities between all user data are calculated individually to select the account with the highest similarity of F_A, F_B for matching.

Step 3: Pruning filter: In order to obtain the best match, reverse confirmation is needed to ensure the precision of the result. F_A is used as the source account in the *Step 1*. Here, F_B is used as the source account for reverse acknowledgment. If the matching result is consistent, then user accounts are retained; otherwise, user accounts are discarded. The whole process of user identification is summarized in Algorithm 2.

Algorithm 2: User Identification Algorithm

Input: Source network account user data information vector F_A, user data information vector $(F_i)_{i=1}^n$ of all accounts in the target network. Candidate set filter condition C corresponding filter user data C_p, Filter similarity matching threshold T_c, account similarity matching threshold T_s.
Output: Account F_B matching F_A.

```
foreach $F_i$ in $(F_i)_{i=1}^n$
  Calculate the similarity of $F_A$ and $F_i$ single user data $C_p$ according to condition C
    $v_p^{ai} = SimFunc(m_p^a, m_p^i)$
    if $v_p^{ai} \geq T_c$
      $c++$
        Add $F_i$ to matching candidate set $(CF_i)_{i=1}^c$
end
  max=0
foreach $CF_i$ in $(CF_i)_{i=1}^c$
  Calculate user similarity using Equation 10
    $Sim(F_A, CF_i)$
    if $Sim_w(F_A, CF_i) > max$
      $max = Sim_w(F_A, CF_i)$, $d = i$
end
    if $max > T_s$
return $F_B$
else return null
```

IV. ANALYSIS OF EXPERIMENTAL RESULTS

In order to verify the effectiveness of the proposed algorithm in this paper. All the experiments were conducted in the computer with 8G memory and 3.9GHz CPU. The ratio between the training set and the test set is set as 3:1. Since the missing data and the noise data that obviously do not match the data type still exist in the user data, the relevant filling method can be adopted to supplement and replace. User profile data and user-generated data from two major social networks, Facebook and Twitter, are selected for cross-social network user identification. Five open datasets of foreign mainstream social networks are provided [29]. Precision rate, recall rate, as well as the F-measure (F1) are used as evaluation criteria. The formulae are expressed as:

$$Precision = \frac{tp}{(tp + fp)} \tag{11}$$

$$Recall = \frac{tp}{(tp + fn)} \tag{12}$$

$$F1 = 2\frac{precision \times recall}{(precision + recall)} \tag{13}$$

where *tp* denotes the number of the same users that are correctly matched, *fp* denotes the number of users that are matched but are not the same, and *fn* denotes the number of users that are not matched but are the same users.

A. THE IMPACT OF USER-GENERATED DATA ON MATCHING RESULTS

The user behavior analysis algorithm based on frequent pattern mining calculates the similarity by cosine similarity and variance. The effectiveness of the proposed algorithm can be seen from the experimental results.

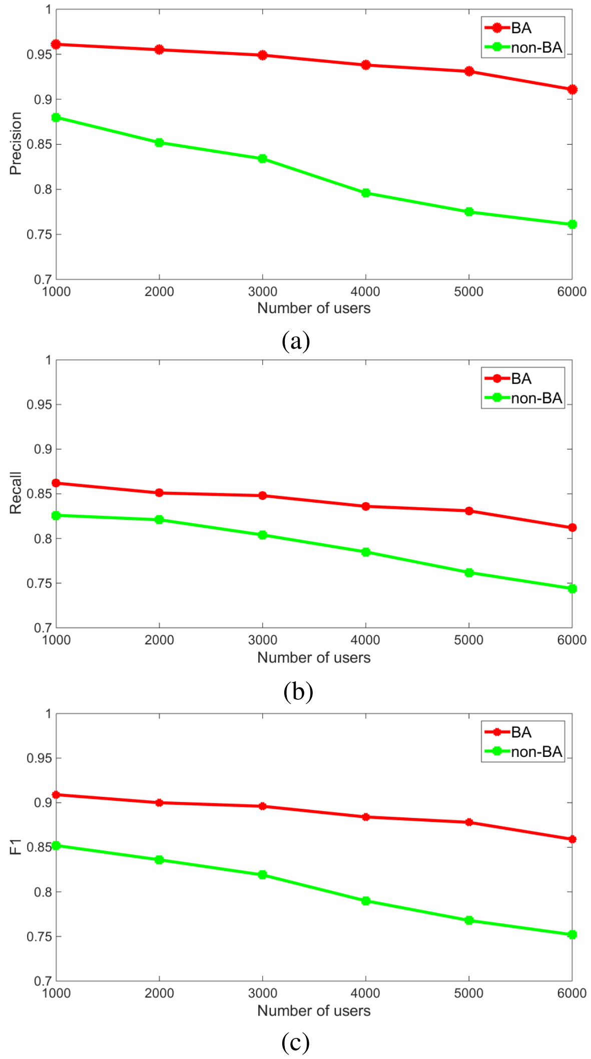
TABLE 2. Comparison of BA and non-BA evaluation indicators.

Number of users	BA			non-BA		
	Precision	Recall	F1	Precision	Recall	F1
1000	0. 961	0. 862	0. 909	0. 880	0. 826	0. 852
2000	0. 955	0. 851	0. 900	0. 852	0. 821	0. 836
3000	0. 949	0. 848	0. 896	0. 834	0. 804	0. 819
4000	0. 938	0. 836	0. 884	0. 796	0. 785	0. 790
5000	0. 931	0. 831	0. 878	0. 775	0. 762	0. 768
6000	0. 911	0. 812	0. 859	0. 761	0. 744	0. 752

Table 2 and Fig. 5 compare the proposed algorithm in terms of precision rate, recall rate, as well as the F1. Some user data are processed (BA) and other user data are not processed (non-BA). BA refers to that processing the user data involves using the proposed algorithm to mine and analyze the user-generated data. Non-BA refers to that user-generated data is not used. In other words, only user profile data is used for cross-social network user identification analysis. Due to the increasing number of users, the similarity of user data is also increasing, which ultimately impacts the matching results. Although the results in Fig. 5 indicate that both BA and non-BA show a downward trend, the decline rate in BA is lower than that in non-BA, which also indicates the effectiveness of the proposed algorithm.

B. THE EFFECT OF WEIGHT ALLOCATION ON THE MATCHING RESULTS

The posterior probability-based information entropy weight allocation algorithm is used to assign the weights of user data with the aim of improving precision. The proposed algorithm also considers the characteristics and practical meanings of different user data. Moreover, the algorithm avoids the shortcomings of the subjective weighting method, and assigns weights to the user data by combining the information contained in the user data with the posterior probability. Therefore, the weight credibility corresponding to user data is more reasonable.



(a)

(b)

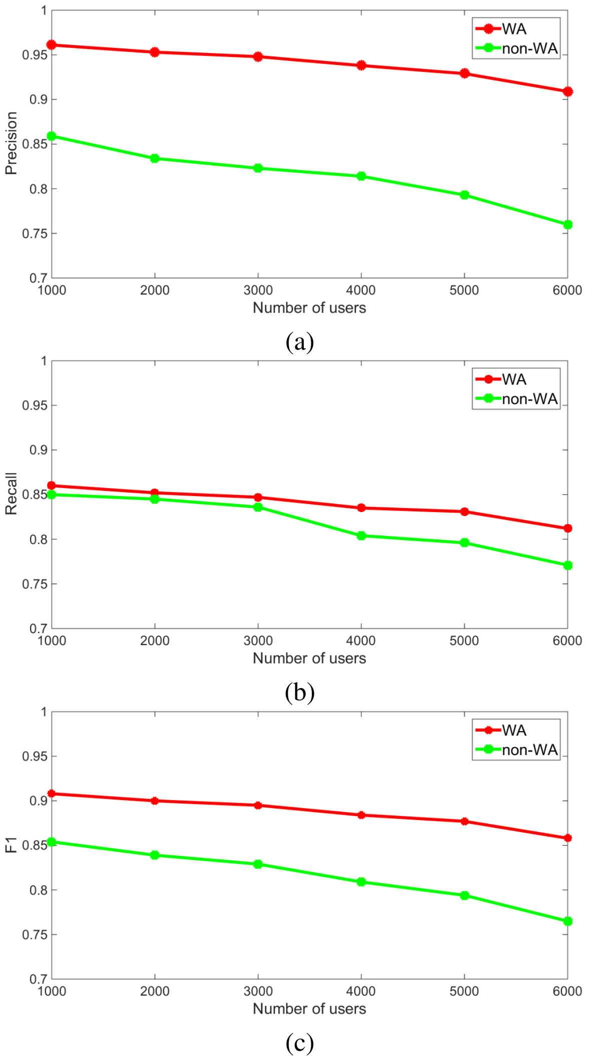
(c)

FIGURE 5. Comparison of BA and non-BA. (a) Precision comparison. (b) Recall comparision. (c) F1 comparison.

TABLE 3. Comparison of IW and non-IW evaluation indicators.

Number of users	IW			non-IW		
	Precision	Recall	F1	Precision	Recall	F1
1000	0. 961	0. 860	0. 908	0. 859	0. 850	0. 854
2000	0. 953	0. 852	0. 900	0. 834	0. 845	0. 839
3000	0. 948	0. 847	0. 895	0. 823	0. 836	0. 829
4000	0. 938	0. 835	0. 884	0. 814	0. 804	0. 809
5000	0. 929	0. 831	0. 877	0. 793	0. 796	0. 794
6000	0. 909	0. 812	0. 858	0. 760	0. 771	0. 765

Table 3 and Fig. 6 illustrate the effectiveness of the proposed algorithm from three perspectives: precision rate, recall rate, as well as the F1. Concerning the problem of weight allocation (i.e., whether to adopt the posterior



(a)

(b)

(c)

FIGURE 6. Comparison of IW and non-IW. (a) Precision comparison. (b) Recall comparision. (c) F1 comparison.

probability information entropy weight allocation algorithm), the user data is processed (IW) and not processed (non-IW) respectively.

As shown in Fig. 6, it is clear that use the proposed algorithm than using data without the posterior probability information entropy weight allocation algorithm in terms of precision rate, recall rate, as well as the F1. This is because the proposed algorithm can reasonably allocate the corresponding weights according to the importance of user data. Compared with empirical weight allocation, the proposed algorithm is more robust. Due to the negative correlation between the number of users and the evaluation index, i.e., when the number of users keeps increasing, there may be a high similarity but no match for the user, which has an impact on the final results, resulting in a downward trend in the evaluation index. Since we have adopted the proposed algorithm for user data, the decline rate is lower than that for unprocessed user data.

V. CONCLUSION

In view of the differences in user behavior, we propose a personalized user behavior analysis algorithm based on frequent pattern mining in this paper. The user data weight is optimized by the posterior probability-based information entropy weight allocation algorithm. The user data weight is not only dependent on the information entropy, but is also based on the prior probability. Therefore, the weight of each user data is more reliable. The experimental results show that the user behavior data is analyzed, which improves the precision rate, recall rate, as well as the F1 of user's matching results.

REFERENCES

[1] X. Zhou, X. Liang, H. Zhang, and Y. Ma, "Cross-platform identification of anonymous identical users in multiple social media networks," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 2, pp. 411–424, Feb. 2016.

[2] M. M. Mostafa, "More than words: Social networks' text mining for consumer brand sentiments," *Expert Syst. Appl.*, vol. 40, no. 10, pp. 4241–4251, 2013.

[3] D. Perito, C. Castelluccia, M. A. Kaafar, and P. Manils, "How unique and traceable are usernames?" in *Proc. Int. Symp. Privacy Enhancing Technol. (PETS)*, 2011, pp. 1–17.

[4] J. Liu, F. Zhang, X. Song, Y.-I. Song, C.-Y. Lin, and H.-W. Hon, "What's in a name?: An unsupervised approach to link users across communities," in *Proc. 6th ACM Int. Conf. Web Search Data Mining*, 2013, pp. 495–504.

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

[6] O. Goga, D. Perito, H. Lei, R. Teixeira, and R. Sommer, "Large-scale correlation of accounts across social networks," Univ. California Berkeley, Berkeley, CA, USA, Tech. Rep. TR-13-002, 2013.

[7] N. Ye, L. Zhao, L. Dong, G. Bian, E. Liu, and G. J. Clapworthy, "User identification based on multiple attribute decision making in social networks," *China Commun.*, vol. 10, no. 12, pp. 37–49, 2013.

[8] P. Jain, P. Kumaraguru, and A. Joshi, "@ i seek'fb. me': Identifying users across multiple online social networks," in *Proc. 22nd Int. Conf. World Wide Web Companion*, 2013, pp. 1259–1268.

[9] P. Jain and P. Kumaraguru. (2012). "Finding nemo: Searching and resolving identities of users across online social networks." [Online]. Available: https://arxiv.org/abs/1212.6147

[10] M. Almishari and G. Tsudik, "Exploring linkability of user reviews," in *Proc. Eur. Symp. Res. Comput. Secur. (ESORICS)*, 2012, pp. 307–324.

[11] X. Kong, J. Zhang, and P. S. Yu, "Inferring anchor links across multiple heterogeneous social networks," in *Proc. 22nd ACM Int. Conf. Inf. Knowl. Manage.*, 2013, pp. 179–188.

[12] O. Goga, H. Lei, S. H. K. Parthasarathi, G. Friedland, R. Sommer, and R. Teixeira, "Exploiting innocuous activity for correlating users across sites," in *Proc. 22nd Int. Conf. World Wide Web*, 2013, pp. 447–458.

[13] X. H. Han, L. H. Wang, S. J. Xu, G. Q. Liu, and D. W. Zhao, "Linking social network accounts by modeling user spatiotemporal habits," in *Proc. IEEE Int. Conf. Intell. Secur. Inform.*, Jul. 2017, pp. 19–24.

[14] Y. Li, Z. Zhang, Y. Peng, H. Yin, and Q. Xu, "Matching user accounts based on user generated content across social networks," *Future Gener. Comput. Syst.*, vol. 83, pp. 104–115, Jun. 2018.

[15] J. Haupt, B. Bender, B. Fabian, and S. Lessmann, "Robust identification of email tracking: A machine learning approach," *Eur. J. Oper. Res.*, vol. 271, no. 1, pp. 341–356, 2018.

[16] S. Bartunov, A. Korshunov, S. T. Park, W. Ryu, and H. Lee, "Joint link-attribute user identity resolution in online social networks," in *Proc. 6th SNA-KDD Workshop*, 2012, pp. 1–9.

[17] X. Zhou, X. Liang, X. Du, and J. Zhao, "Structure based user identification across social networks," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 6, pp. 1178–1191, Jun. 2018.

[18] L. Xing, Q. Ma, and L. Jiang, "Microblog user recommendation based on particle swarm optimization," *China Commun.*, vol. 14, no. 5, pp. 134–144, 2017.

[19] E. Raad, R. Chbeir, and A. Dipanda, "User profile matching in social networks," in *Proc. 13th Int. Conf. Netw.-Based Inf. Syst. (NBiS)*, 2010, pp. 297–304.

[20] K. Cortis, S. Scerri, I. Rivera, and S. Handschuh, "An ontology-based technique for online profile resolution," in *Social Informatics*. Berlin, Germany: Springer, 2013, pp. 284–298.

[21] T. Iofciu, P. Fankhauser, F. Abel, and K. Bischoff, "Identifying users across social tagging systems," in *Proc. 5th Int. AAAI Conf. Weblogs Social Media*, 2011, pp. 522–525.

[22] F. Abel, E. Herder, G.-J. Houben, N. Henze, and D. Krause, "Cross-system user modeling and personalization on the social Web," *User Model. User-Adapt. Interact.*, vol. 23, nos. 2–3, pp. 169–209, 2013.

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

[24] A. Esfandyari, M. Zignani, S. Gaito, and G. P. Rossi, "User identification across online social networks in practice: Pitfalls and solutions," *J. Inf. Sci.*, vol. 44, no. 3, pp. 377–391, 2016.

[25] Y. Li, Y. Peng, Z. Zhang, H. Yin, and Q. Xu, "Matching user accounts across social networks based on username and display name," in *Proc. World Wide Web*, 2018, pp. 1–23.

[26] H. V. Nguyen and L. Bai, "Cosine similarity metric learning for face verification," in *Proc. Asian Conf. Comput. Vis.* New York, NY, USA: Springer-Verlag, 2010, pp. 709–720.

[27] G. Wang, X. Zhang, S. Tang, H. Zheng, and B. Y. Zhao, "Unsupervised clickstream clustering for user behavior analysis," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2016, pp. 225–236.

[28] K. A. Alam, R. Ahmad, and K. Ko, "Enabling far-edge analytics: Performance profiling of frequent pattern mining algorithms," *IEEE Access*, vol. 5, pp. 8236–8249, 2017.

[29] M. Yan, J. Sang, and C. Xu, "Unified YouTube video recommendation via cross-network collaboration," in *Proc. 5th ACM Int. Conf. Multimedia Retr.*, New York, NY, USA, 2015, pp. 19–26.

KAIKAI DENG received the B.S. degree in electronic information engineering from the Anyang Institute of Technology, China, in 2017. He is currently pursuing the M.S. degree in information and communication engineering with the Henan University of Science and Technology, China. His research interests include data mining, social media trust analysis, and social computing.

LING XING received the B.S. degree in electronic engineering from the Southwest University of Science and Technology, China, in 2002, the M.S. degree in electronic engineering from the University of Science and Technology of China, in 2005, and the Ph.D. degree in communication and information system from the Beijing Institute of Technology, in 2008.

In 2007, she was a Visiting Scholar with the Illinois Institute of Technology, Chicago, USA. She is currently a Professor with the School of Information Engineering, Henan University of Science and Technology, China. Her research interests include multimedia semantic mining, private preserving, and social computing.

LONGSHUI ZHENG received the B.S. degree in electronic science and technology from Shangqiu Normal University, China, in 2014, and the M.S. degree in communication and information system from the Southwest University of Science and Technology, China, in 2017. His research interests include social media trust analysis and big data mining.

HONGHAI WU received the B.S. degree from Zhengzhou University, China, in 2001, and the M.S. and Ph.D. degrees from the Beijing University of Posts and Telecommunications, China, in 2007 and 2015, respectively. He was with China United Telecommunications Co., Ltd., from 2007 to 2011. He is currently with the College of Information Engineering, Henan University of Science and Technology, Luoyang, China. His research interests include delay/disrupted tolerant networks, opportunistic networks, and video delivery.

PING XIE was born in Hunan, China, in 1984. She received the B.Sc. degree in communication engineering and the M.S. degree in communication and information systems from the Kunming University of Science and Technology, Kunming, China, in 2007 and 2011, respectively, and the Ph.D. degree from the Beijing University of Posts and Telecommunications, China, in 2014. She is currently a Lecturer with the Henan University of Science and Technology. Her research interests include cognitive radio networks, physical layer security, and resource allocation for fading channels.

FEIFEI GAO (M'09–SM'14) received the B.Eng. degree from Xi'an Jiaotong University, Xi'an, China, in 2002, the M.Sc. degree from McMaster University, Hamilton, ON, Canada, in 2004, and the Ph.D. degree from the National University of Singapore, Singapore, in 2007.

He was a Research Fellow with the Institute for Infocomm Research (I2R), A*STAR, Singapore, in 2008, and an Assistant Professor with the School of Engineering and Science, Jacobs University, Bremen, Germany, from 2009 to 2010. In 2011, he joined the Department of Automation, Tsinghua University, Beijing, China, where he is currently an Associate Professor. His research interests include communication theory, signal processing for communications, array signal processing, and convex optimizations, with particular interests in MIMO techniques, multi-carrier communications, cooperative communication, and cognitive radio networks. He has authored/coauthored more than 100 refereed IEEE journal papers and more than 100 IEEE conference proceeding papers, which have been cited more than 4000 times from Google Scholar.

Dr. Gao has served as a Symposium Co-Chair for the 2014 IEEE Vehicular Technology Conference Fall, the 2014 IEEE Global Communications Conference, and the 2015 IEEE Conference on Communications, and a Technical Committee Member for many other IEEE conferences. He has also served as an Editor for the IEEE Transactions on Wireless Communications, the IEEE Communications Letters, the IEEE Signal Processing Letters, the IEEE Wireless Communications Letters, the *International Journal on Antennas and Propagations*, and *China Communications*.