A user identity matching method based on integrating account attributes

Y Z Yang, H T Yu, R Y Huang
National Digital Switching System Engineering & Technological R&D Center, Zhengzhou, China
yizhuo_yang@foxmail.com

Abstract: Aiming at the low utilization rate of attribute information and the lack of mining of the correlation among attributes of the existing cross-social network user identity matching algorithms, we proposed an algorithm for user identity matching across social networks utilizing fuzzy measure and Choquet integrals. Firstly, according to the characteristics of different attributes, we determined different similarity calculation strategies; Secondly, we utilized particle swarm optimization method to calculate the fuzzy density of each attribute; Then Choquet integral was utilized to calculate the similarity of two accounts; Finally, the similarity was compared with the preset matching threshold and the final matching result was obtained. The experimental results in multiple sets of data showed that the average F1 value of the proposed algorithm reaching 84.5%. The performance is not only better than traditional machine learning methods, but also better than several baseline algorithms. It can be more accurate to identify the same user’s accounts in multiple social networks according to the attribute information.

1. Introduction
With the continuous development of internet technology, the online social network (OSN) has gradually become an integral part of people's daily life. Due to the different functions and target users of various social networks, a single social network usually fails to satisfy all of the user’s online social needs. So, many users register their accounts in multiple social networks at the same time. It is reported that 42% of online users are using multiple social networks, and 93% Instagram users have accounts in Facebook [1]. User identity matching across social network is to identifying the same user’s account registered in multiple online social networks according to the various online social network data. These algorithms have great practical significance and utilization value in many fields. First, in the commercial field, the accuracy of the product recommendation system can be improved a lot and considerable economic benefits can be brought about. Second, in the cyber security field, various malicious behaviors due to the anonymity of social networks can be greatly overcome. There are many false information, rumors and inappropriate speech. It will spend a lot to ban these malicious acts directly. If the technique of user identity matching can be utilized to link multiple malicious accounts registered by one malicious user, the cyber security will be greatly maintained. What’s more, user identity matching can help a lot in cyber friend recommendation and resolving, the cold start problem, et al.

2. Background and related works
At present, most of the researches on user identity matching across social networks adopted three kinds of data: user generated content(UGC), network topology and account attributes. The UGC based...
methods achieve user identity matching by analyzing the contents (texts, pictures, videos, etc.) with time and place. Some researches utilized the contents posted by a user to analyze his/her interest, based on which achieve the matching[2]. The network topology based methods achieve the matching by the linkages between users. Some researchers proposed to use the network embedding method to mine the topology structure in social networks[3]. But there are disadvantages of methods based on UGC and network topology, that is, the dependence on data is too high. The UGC based methods are difficult to achieve the matching when there are few contents. At the same time, network topology based methods is bound to have a large number of nodes with similar structural information when the data is large, making it difficult to accurately determine the matching account. The account attributes based methods analyze the similarity of basic attributes such as username, age, location and occupation to achieve the matching. Therefore, we utilized account attributes to achieve the matching.

Some researchers utilized naïve Bayes classifier to achieve the matching[4]. Some others comprehensively considered the characteristics of various attributes and used a fusion of variety of classifiers to further enhance the performance[5]. Although the performances of above methods are pretty well, there are still some problems. First, most researchers used the same similarity calculation method for different attributes. Second, these methods didn’t take the relationships of attributes into consideration. According to these problems, we proposed a Choquet integral based user identity matching method. First, we used different ways to calculate similarities between different attributes. Then, we utilized the fuzzy measure and Choquet integral to integrate these criteria to match the user identity.

3. Method Description

3.1. Fuzzy measure and Choquet integral

Sugeno created the fuzzy integral theory, which gave a fuzzy measure on an ordinary measurable set and used the fuzzy measure to define the Sugeno fuzzy integral[6]. Choquet integral is also a kind of fuzzy integral and is usually utilized in multicriteria decision making[7]. We regarded each attribute similarity a criterion and utilized Choquet integral to integrate these criteria to match the user identity.

3.1.1. Fuzzy measure.

We use two definitions to definite fuzzy measure which is mentioned in this paper:

**Definition1:** A fuzzy measure $g$ on $U$ is a function $g: M \rightarrow [0, 1]$, satisfying the following axioms:

1. $g(\emptyset) = 0, g(U) = 1$
2. $A \subseteq B \Rightarrow g(A) \leq g(B)$
3. $\bigwedge_{i=1}^{n} A_i \Rightarrow \lim_{n \to \infty} g(A_i) = g(A)$ (n $\geq 1$)

where $M \in \mathcal{P}(U)$ and $U = \{u_1, u_2, ..., u_n\}$, $\mathcal{P}(U)$ is the power set of U.

**Definition2:** A $g_i$ fuzzy measure is a fuzzy measure, satisfying the following axioms:

1. $\forall A, B \in \mathcal{P}(U)$ and $A \cap B = \emptyset$, $g_i(A \cup B) = g_i(A) + g_i(B) + \lambda g_i(A)g_i(B)$
2. $\lambda \geq 1 \lambda \prod_{i=1}^{n} (1 + \lambda g_i(u_i))$

where $\lambda \in (-1, +\infty)$, $g_i(u_i)$ is the fuzzy density of criterion $i$, and we used $g_i$ for convenience. It can be proven that for a given set of all $g_i$, there is only one corresponding $\lambda$[8]. We can calculate all fuzzy measures based on the fuzzy density according to Def.2 by:

$$g_i(U) = \frac{1}{\lambda} \prod_{i=1}^{n} (1 + \lambda g_i) - 1$$  \hspace{2cm} (1)
for each $U_i = \{u_{i,1}, u_{i,2}, ..., u_{i,j}\}$ ($i \leq n$).

### 3.1.2 Discrete Choquet integral.

The Discrete Choquet integral is defined by the following definition:

**Definition 3:** The discrete Choquet integral of a $g_i$ fuzzy measure is defined by:

$$C_{\sigma} = \sum_{i=1}^{n} u_{\sigma(i)} \left[ g_i (A_{\sigma(i)}) - g_i (A_{\sigma(i+1)}) \right]$$

(2)

where $\sigma$ indicates that the indices have been permuted so that $u_{\sigma(1)} \leq u_{\sigma(2)} \leq \cdots \leq u_{\sigma(n)}$ and $A_{\sigma(i)} = \{u_{\sigma(i)}, u_{\sigma(i+1)}, \ldots, u_{\sigma(n)}\}$. It can be seen that Choquet integral takes the correspondence between criteria into consideration.

### 3.1.3 PSO based fuzzy density optimization.

The key to utilizing Choquet integral to achieve the matching is the fuzzy density. We transformed the problem of determining the fuzzy density into an optimization process of a series of parameters. The PSO algorithm was utilized to determine the fuzzy density. It is an evolutionary computational technique derived from the study of bird predation behavior. Each solution to the optimization problem is a bird in the search space, called a “particle”. Each particle has a fitness value to measure the distance of the particle and the solution. Particles constantly update their position by changing their direction and speed according to their own experience and the global experience. After the iteration, all particles will gradually tend to the optimal solution.

We first constructed $N$ k-dimensional vectors $\{x_1, x_2, ..., x_k\}$ and each vector represented a particle, each dimension of the vector $x_i$ represented the fuzzy density of the $i$th attribute. We used F1 value as the final evaluation method, so the fitness value here is determined by F1 value. The formulas for the particle updating of its speed and position are defined by:

$$v_{i,j} (t+1) = \omega v_{i,j} (t) + c_1 r_{i,j} (t) (y_{i,j} (t) - x_{i,j} (t)) + c_2 r_{2,j} (t) (\hat{y}_j (t) - x_{i,j} (t))$$

(3)

$$x_i (t+1) = x_i (t) + v_{i,j} (t+1)$$

(4)

where $v_{i,j} (t)$ indicates the speed of the $j$th dimension of the particle $i$, $y_{i,j} (t)$ indicates the value of the $j$th dimension with the max fitness value in last $t$ iteration, $\hat{y}_j (t)$ indicates the value of the $j$th dimension of the best particle in all the $N$ particles. $\omega$ indicates the inertia weight. $c_1$ and $c_2$ are the acceleration constants and $r_{i,j} (t), r_{2,j} (t) \sim U (0, 1)$. In this paper, we made $\omega$ change with the iteration number by:

$$\omega (t) = \omega_{\text{max}} - t \times \left( \frac{\omega_{\text{max}} - \omega_{\text{min}}}{t_{\text{max}}} \right)$$

(5)

where $\omega_{\text{max}} = 0.9$, $\omega_{\text{min}} = 0.1$. $t_{\text{max}} = 1000$ is the maximum of the iteration. $c_1 = c_2 = 2$[9] and the number of particles $N$ is 30[10].

### 3.2 Method of attribute similarity calculation

We utilized different similarity calculation methods for different attributes. The attributes we selected included: username, personal description, location, language and gender. The specific method for calculating the similarity of each attribute is as follows.

1. **username.** Username is the most easily obtained information in the social network. Many researchers have achieved the matching based solely on username attributes. We vectorized the username attribute so that each dimension represented a certain characteristic of the username string. First, we removed the special characters in the username, kept only the numbers and letters. Second, we
made the usernames keep the same letter case. As for the feature extraction, each first occurrence of two consecutive character combinations was used as a feature, and its frequency of occurrence was counted as feature value. Then the frequency-inverse document frequency (TF-IDF) was used to calculate the weight of each feature. The specific calculation method is as follows:

\[
idf_i = \log \frac{|P|}{\sum_{j \in P} \delta_{p_j, p_i}}
\]

(6)

where \( |P| \) is the total number of the users and \( p_i \) is one of the usernames from \( |P| \). Finally, the cosine similarity between two username vectors was calculated as the username attribute similarity. We use three usernames here as an example: lilian, diana and lily. The result of feature extraction is shown in the Table1:

Table 1. An example of username attribute character extraction

|     | li | il | ia | an | di | na | ly |
|-----|----|----|----|----|----|----|----|
| lilian | 2  | 1  | 1  | 1  | 0  | 0  | 0  |
| diana | 0  | 0  | 1  | 1  | 1  | 1  | 0  |
| lily  | 1  | 1  | 0  | 0  | 0  | 0  | 1  |

After calculation, the similarities between lilian and the other two usernames diana and lily are 0.185 and 0.371 respectively. This result is consistent with our subjective feelings and illustrated the rationality of the similarity calculation method for usernames.

(2) Personal description. The personal description of a user is a long text which is not easy to calculate the similarity. We utilized the Latent Dirichlet Allocation (LDA) to obtain the topics described by each text. LDA is a topic model which assigns each word from the text to a topic. We chose some topics including career, interest, specialty, and so on. After obtaining the topic probability distribution of each text, the similarity between two texts \( p \) and \( q \) is as follows:

\[
Sim(p, q) = 1 - JS(\theta_p, \theta_q)
\]

(7)

where \( \theta_p, \theta_q \) are the topic distributions, \( JS(\theta_p, \theta_q) \) is the Jensen-Shannon divergence (JS divergence) of the two topic distributions[11], which is calculated by:

\[
JS(\theta_p, \theta_q) = \frac{1}{2} \left[ KL(\theta_p, \theta_m) + KL(\theta_q, \theta_m) \right]
\]

(8)

where \( \theta_m = (\theta_p + \theta_q)/2 \) and \( KL(\theta_p, \theta_m) = \sum_{k=1}^{|P|} \theta_{p,k} \cdot \log(\theta_{p,k}/\theta_{m,k}) \). KL divergence measures the “distance” between two probability distributions. Because of the asymmetry and unrestricted range, it is difficult to apply the similarity calculation. JS divergence overcomes these shortcomings. We used this to calculate the similarity between two texts and got a pretty good result.

(3) Location. Different social networks have different formats for location information and it is not convenient to use common methods. We utilized the API of the map website to convert the obtained location into latitude and longitude. The similarity between two locations is calculated as follows:

\[
Sim(p_i, p_j) = e^{-\gamma d(p_i, p_j)}
\]

(9)

where \( \gamma \) is the normalization coefficient, \( d(p_i, p_j) \) indicates the distance which is calculated as follows[12]:

\[
d(p_i, p_j) = R \times \arcsin \left[ \cos(\text{Lat}_i) \cdot \cos(\text{Lat}_j) \cdot \cos(\text{Lng}_i - \text{Lng}_j) + \sin(\text{Lat}_i) \cdot \sin(\text{Lat}_j) \right]
\]

(10)

where R indicates the radius of the earth, which is 6371km in this paper.

(4) Language and gender. For these two attributes, similarities don’t make sense. The similarities of these two attributes were only measured by 0 and 1, i.e. if the attribute is the same, the similarity is 1, else the similarity is 0.
3.3 Our method
The method we proposed includes four steps: attribute similarity calculation, fuzzy density optimization and fuzzy measure calculation, Choquet integral calculation, accounts matching. First, we calculated the attribute similarity for each pair of accounts. Second, PSO-based fuzzy density optimization algorithm was utilized to obtain the fuzzy density corresponding to each account attribute. Then, the similarity between each pair of accounts was calculated by (1). Finally, we compared the score with the preset threshold T. In summary, the algorithm proposed in this paper is as follows:

Algorithm 1 Choquet integral based user identity matching method

**Input:** two social networks $G^1 = (\mathcal{N}^1, \mathcal{P}^1)$ and $G^2 = (\mathcal{N}^2, \mathcal{P}^2)$, threshold $T$.

**Output:** The matching set $A$.

1) $A = \emptyset$
2) for each $(n^1_m, n^2_n)$ from $G^1$ and $G^2$
3) calculate $\text{Sim}(n^1_m, n^2_n)$ by Sec.3.2
4) calculate $g_i$ by Sec.3.1.3
5) calculate $\lambda$ by Def.2
6) calculate $g_U(\lambda)$ by eq. (1)
7) for each $(n^1_m, n^2_n)$ from $G^1$ and $G^2$
8) calculate the score of similarity by eq. (2)
9) if score $> T$
10) join the pair of accounts to $A$
11) return $A$

4. Experiments and results

4.1 Datasets
We selected Google+, Facebook and Twitter which currently have a large number of users to verify the performance of our method. Google+ allows users to add their own links to other social networks on the homepage, which is intended to allow users to extend their friendships to more social networks[13]. We collected links to Facebook and Twitter through Google+ to collect user account attribute information. After filtering the invalid or empty links, 5970 Google+ accounts were eventually acquired, while 4376 Facebook accounts and 5632 Twitter accounts were acquired at the same time. According to these accounts, we obtained three datasets for the experiments.

4.2 Evaluation Method
We utilized precision, recall and F1 value which are usually used in classification problems as the evaluation index of our experiments. The precision $P$ represents the rate of true positive and all positive results. The recall $R$ represents the rate of true positive and all true results. The F1 value comprehensively considers the two indicators. At the same time, F1 value is the fitness value of the PSO algorithm in our paper. The calculation method is as follows:

$$F1 = \frac{2PR}{P + R}$$
4.3. Experimental results

4.3.1 Fuzzy density optimization.

We selected username, personal description, location, language and gender as user attributes. The parameter setting of the PSO algorithm is described in Sec. 3.1.3. According to the parameter value, the fuzzy density of different attributes in different are shown in Table 2.

|                     | Username | Personal Description | Location | Language | Gender |
|---------------------|----------|----------------------|----------|----------|--------|
| Facebook-Google+    | 0.7634   | 0.2641               | 0.2146   | 0.1096   | 0.0235 |
| Twitter-Google+     | 0.8041   | 0.3062               | 0.1167   | 0.0806   | 0.0973 |
| Twitter-Facebook    | 0.7705   | 0.2267               | 0.1498   | 0.0764   | 0.0632 |

As can be seen from the table, the fuzzy density of the user name attribute is the largest no matter in which dataset. It showed that the username was the most important attribute during the matching. The fuzzy density of personal description and location has little difference, and they can affect the similarity score to some extent. In contrast, the influence of the two attributes of language and gender on the final result is relatively limited because there are many users have the same attribute in a social network.

4.3.2 Performance comparison

The baseline algorithms we chose in the experiment included SVM algorithm, Goga’13 algorithm, Li’17 algorithm and MOBIUS algorithm. Goga’13 firstly utilized various machine learning algorithms to analyze the ability of each attribute in the matching. Then the attribute weights were assigned according to the ability. Li’17 proposed several similarity indicators for the characteristics of the username attribute and combined them to better identify two usernames belonging to the same user. MOBIUS performed an analysis of each attribute information of the accounts according to human behaviors, Markov chain was used to predict string attributes and a naïve Bayes classifier was used to fuse and match multiple attributes. The experimental results are shown in Figure 1-3.

Experimental results showed that the proposed method had better performance than several baseline methods. The SVM algorithm is a data-sensitive algorithm. When there were missing values, the matching effect will be greatly affected. Although Li’17 had a high precision rate in every dataset, it’s recall rate is too low and the performance was affected. This is because there are some users with totally different usernames in different social networks. Due to the lack of other attributes, the F1 values of Li’17 were lower. As for Goga’13 and MOBIUS, they considered a variety of attributes and evaluated the importance of each attribute. But the relationships of these attributes are not considered into during the matching. Each attribute was regarded as independent, and the influence of multiple attributes was ignored. We made use of fuzzy measure theory and Choquet integral theory to achieve the matching. Our method not only considered the importance of each attribute, but also considered the influence of multiple attributes. The experimental results showed that the average precision reached 87.3%, the average recall reached 81.9%, and the average F1 value reached 84.5%. The F1 value exceeded several baseline methods, which verified the effectiveness of the proposed algorithm.
5. Conclusion and future work
In this paper, we proposed a cross-social networks user identity matching method based on account attributes. First, the fuzzy density of each attribute was obtained by the PSO algorithm. Then we calculated the fuzzy measures as the influence of multiple attributes. Finally, the similarity of two accounts was calculated by Choquet integrals. Experiments were performed in three datasets consisting of Google+, Twitter and Facebook, which verified the effectiveness of the proposed method.

Due to the lack of an algorithm to filter out the totally unmatched accounts, the number of accounts to be matched was too large, which affected the efficiency of the algorithm. At the same time, due to the lack of processing of the missing account attribute scenario, the method needs datasets with high quality. Our next work will solve the above two problems, design a filtering algorithm to reduce the calculation cost, and at the same time design a processing method for missing attributes so that the method can be more robust.

References
[1] Shu K, Wang S, Tang J, Zafarani R and Liu H 2017 User identity linkage across online social networks: a review ACM SIGKDD Explorations Newsletter. 18(2) 5-17
[2] Nie Y, Jia Y, Li S, Zhu X, Li A and Zhou B 2016 Identifying users across social networks based on dynamic core interests Neurocomputing. 210 107-115
[3] Man T, Shen H, Liu S, Jin X and Cheng X 2016 Predict anchor links across social networks via an embedding approach International Joint Conference on Artificial Intelligence. 1823-29
[4] Zafarani R and Liu H 2013 Connecting users across social media sites: a behavioral-modeling approach. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 41-49
[5] Goga O, Perito D, Lei H, Teixeira R and Sommer R 2013 Large-scale correlation of accounts across social networks University of California at Berkeley, Berkeley, California, Tech. Rep. TR-13-002
[6] Sugeno M. 1974 Theory of fuzzy integrals and its applications Doctoral Thesis. Tokyo Institute of Technology
[7] Grabisch M and Roubens M 2000 Application of the choquet integral in multicriteria decision making, Fuzzy Measures and Integrals-Theory and Applications. 348-374
[8] Yang L, Gao Y and Ling W X 2011 Principle and application of fuzzy mathematics(Guangzhou: South China University of Technology Press) p 217
[9] Brits R, Engelbrecht A P and Van den Bergh F 2002 A niching particle swarm optimizer Proceedings of the 4th Asia-Pacific conference on simulated evolution and learning. 2 692-696
[10] Shi Y and Eberhart R C 1998 Parameter selection in particle swarm optimization International Conference on Evolutionary Programming. 591-600
[11] Lin J 1991 Divergence Measures based on the shannon entropy IEEE Transactions on Information Theory. 37(1) 145-151
[12] Fu Z 1990 The calculation of great-circle distance and great circle route Journal of China Civil Aviation University. 8(1) 145-151
[13] Yan M, Sang J and Xu C 2015 Unified youtube video recommendation via cross-network collaboration. Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. 19-26