A Solution to Tweet-Based User Identification Across Online Social Networks

Yongjun Li(✉), Zhen Zhang, and You Peng

School of Computer, Northwestern Polytechnical University,
Xiàn, Shaanxi 710072, China
lyj@nwpu.edu.cn

Abstract. User identification can help us build better users' profiles and benefit many applications. It has attracted many scholars' attention. The existing works with good performance are mainly based on the rich online data. However, due to the privacy settings, it is costless or even difficult to obtain the rich data. Besides some profile attributes do not require exclusivity and are easily faked by users for different purposes. This makes the existing schemes are quite fragile. Users often publicly publish their activities on different social networks. This provides a way to overcome the above problem. We aim to address the user identification only based on users' tweets. We first formulate the user identification based on tweets and propose a tweet-based user identification model. Then a supervised machine learning based solution is presented. It consists of three key steps: first, we propose several algorithms to measure the spatial similarity, temporal similarity and content similarity of two tweets; second, we extract the spatial, temporal and content features to exploit information redundancies; Afterwards, we employ the machine learning method for user identification. The experiment shows that the proposed solution can provide excellent performance with F1 values reaching 89.79%, 86.78% and 86.24% on three ground truth datasets, respectively. This work shows the possibility of user identification with easily accessible and not easily impersonated online data.

Keywords: User identification · Tweet · Social network · Machine learning · Online behavior analysis

1 Introduction

In the last decade, many types of social networking sites have emerged and grown rapidly in Monthly Active Users(MAU). As of April 2017, Twitter has more than 319 million MAUs, and Facebook has 1,968 million MAUs. Sina Microblog has also more than 313 million MAUs [1]. These social sites have changed the way we interact with each other, and make it simple to stay connected in our lives.

Due to the differences in the services provided by online social networks (OSNs), people tend to use different OSNs for different purposes. As we may expect, a user's activities and connections are scattered into several different sites. If we integrate these sites, his better and more complete profile can be

G. Cong et al. (Eds.): ADMA 2017, LNAI 10604, pp. 257–269, 2017.
https://doi.org/10.1007/978-3-319-69179-4_18

built to improve online services, such as community discovery, recommendation, and information diffusion.

To integrate these OSNs, it is necessary to identify users across sites. There are some existing works discussing possible solutions to this problem. Many existing works addressed this problem based on the rich user profile attributes, including screen name, birthday, hometown, etc. [3–10]. Owing to privacy settings, it is high costless or even difficult to obtain the above attributes. On the other hand, these attributes are easily faked by users for different purposes. These limitation make the existing schemes quite fragile [2]. Some researchers leveraged the friend network to identify users [2,11–21]. Taking into account personal privacy, most of users do not make their friend network public. Even if we can obtain the user's friend network, these connections are also sparse. The existing methods based on friend network are also plagued by the above limitations [2]. Some researchers also employed user tweets to identify users based on posting time, location and writing style [22–25]. However, in existing works, the tweets are always used with profiles or friend network together, so these solutions face similar problems as described above.

The tweets posted on different sites by the same user usually contain rich information redundancies. Meanwhile, users often make some of their tweets public and easily accessible. Intuitively, we can identify users solely based on the users' tweets, and break through these limitations. However, the tweets-based method is surely very challenging. The first challenge is that writing style, usually used in the existing works, is difficult to extract from short tweets. The number of tweets the user posting publicly on different sites are serious imbalance. In this study, we calculate semantic similarity of tweets rather than writing style. On the other side, we consider the similarity of any two tweets from different sites to overcome the problem of imbalance. We present a novel framework to tackle the user identification. This method could be applied jointly with other feature-based algorithms for more accurate results.

The rest of the paper is organized as follows. We first introduce the related works in Sect. 2. Then in Sect. 3, we describe the preliminary concepts, and give the problem formulation. In Sect. 4, we present the solution framework and tweet-based user identification across OSNs. Then Sect. 5 shows the experiment results on social networks. In Sect. 6, we conclude the paper.

2 Related Works

In OSN, a user usually creates an identity and constitutes its three major dimensions namely Profile, Content and Network. Each dimension is composed of a set of attributes which describes her and differentiates her from others [26]. The existing works are mainly based on these three dimensions or the hybrid dimensions.

In some existing works, the researchers presented methods which solely use profile attributes to identify users across sites. Liu et al. [3] matched user accounts in an unsupervised approach using usernames. Zafarani et al. [4] presented a MOBIUS method to identify users based on the naming patterns of

usernames. Perito et al. [27] introduced the idea of using username to match the accounts of a user across sites. Liu et al. [29] analyzed usernames' characteristics including length, special character, numeric character etc., and proposed a weighting function of user identification. However, username are not always available, and even in some situation, the username is a numeric string automatically assigned by sites. This makes these existing schemes fragile. Motoyama et al. [6] extended the profile attribute set, and used name, city, school, location, age, email etc. to match user accounts. Iofciu et al. [5] used the similarity between users' profiles to identify users. Abel et al. [7] aggregated user profiles and matched users across systems. Raad et al. [8] addressed the user identification by providing a matching framework based on all the prole's attributes. The proposed framework allowed users to give more importance to some attributes and assign each attribute a different similarity measure. The hybrid methods concluded that user accounts could be accurately matched based on a set of attributes. However, the profile attributes are not exclusive and easily faked by users for different purposes.

Some existing works studied the user identification problem solely based on user network. Zhou et al. [2] proposed a friend relationship-based user identification algorithm. It calculates a match degree for all candidate user matched pairs, and only pairs with top ranks are considered as identical users. Narayanan et al. [19] solely used network structure to analyze privacy and anonymity, which is closely related to user identification issue. Korula et al. [21] presented a mapping algorithm based on the degrees of unmapped users and the number of common neighbors, using two control parameters to finetune performance. Owing to the privacy setting, in many cases, the users' friend networks are not public and accessible across sites. Researchers attempted hybrid approaches to solve this issue. Bartunov et al. [20] considered both the profile and friend network, and proposed an approach based on conditional random fields to identify users. Bennacer et al. [30] also used the friend network and the publicly available profile to iteratively match profiles across OSNs. Malhotra et al. [31] used user profile and friend network to generate the user's digital footprints, and applied automated classifiers for user identification. The above studies show that the friend network has forceful and robust features for user identification. However, this information is often sparse, because only a small portion of users are willing to make their friend network public.

A set of researchers used the content dimension for user identification. Goga et al. [23] used three attributes extracted from the content, timestamp, location and description, to identify users. Kong et al. [22] considered the content and social relationship to solve this issue, and proposed Multi-Network Anchoring to match user accounts. They calculated the combined similarities of user's social, spatial, temporal and text information, and employed SVM classifier to identify users. Almishari et al. [32] studied likability of community-based reviews and show that a high percentage of ostensibly anonymous reviews can be accurately linked to their authors. This study focuses on one single popular site(Yelp). Besides, Jiang et al. [28] assume multiple accounts belonging to the same person

contain the same or similar camera fingerprint information, and identify the user by matching his cameras. In these existing content-based works, only Goga et al. [23] solely used the content to identify users across OSNs. At this point, this works is the same as our work, and we also find that the location of tweets is the most powerful feature to match accounts. Our work also focuses on spatial, temporal and text information extracted from tweets, but it is different from Goga et al.'s work in the information processing.

3 Problem Formulation

In this paper, we focus on studying the tweet-based user identification problem across OSNs. The task of tweet-based user identification is to predict whether a pair of user accounts from two OSNs belongs to the same individual. This problem can easily be generalized to the cases with more than two OSNs.

Suppose we are given an OSN $G^1 = \{V^1, E^1\}$, where V^1 is a set of nodes and E^1 is a set of links. For node $v_i^1 \in V^1$, it represents an offline individual. This individual has a unique account u_i^1 on G^1, and also posts some short public tweets on his page. These tweets are denoted as TW_i^1. The k^{th} tweet of v_i^1 is denoted by $tw_{ik}^1 \in TW_i^1$. The tweet tw_{ik}^1 is triple $tuple(t_{ik}^1, l_{ik}^1, w_{ik}^1)$, where t_{ik}^1 is posting time of tweet tw_{ik}^1, l_{ik}^1 is the posting location or place, and w_{ik}^1 is the set of words that user has used in tweet tw_{ik}^1. Similarly, we define another network as $G^2 = (V^2, E^2)$. u_i^2 denotes the account of $v_i^2 \in V^2$ on G^2. TW_i^2 denotes his tweets on G^2 and tw_{ik}^2 denotes the k^{th} tweet.

The tweets posted by a user on different OSNs provide rich information redundancies and can help identify users across sites. When considering tweet-based user identification, we first need to analyze and measure these information redundancies and solve the following general problem based on analysis results.

Given two tweet sets TW_i^1 and TW_k from two different OSNs, do they belong to the same offline individual?

Tweet-Based User Identification. Suppose we have two OSNs G^1 and G^2, with a small set of identified users across two OSNs, $A = (v_i^1, v_j^2), v_i^1 \in V^1, v_j^2 \in V^2$. $\forall (v_i^1, v_j^2) \in A$, we also know the tweet sets TW_i^1 and TW_j^2. Given two tweet sets TW_m^1 and TW_n^2, where $(v_m^1, v_n^2) \notin A$, the task of tweet-based user identification is to determine whether node v_m^1 and v_n^2 belong to the same individual.

The key issue of tweet-based user identification is to learn a identification function of two user accounts. The main difference from the existing works is that we identify users solely based on users' public tweets. This suggests that we should analyze information redundancies of tweets accurately.

4 Model and Solution Framework

For tweet-based user identification, we have the following basic intuition. If two users post several similar tweets, including similar content, similar posting time or similar location, their two accounts belong to the same individual with high probability.

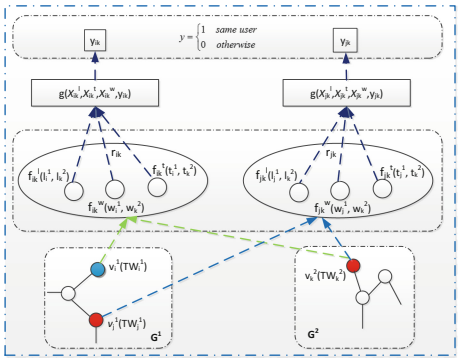


Fig. 1. Graphical representation of T-UIM

4.1 Tweet-Based User Identification Model

Based on the above intuitions, we propose a tweet-based user identification model (T-UIM). Figure 1 shows the graphical representation of the T-UIM. A pair of $users(v_i^1, v_k^2)$ from two OSNs G^1 and G^2 is mapped into a node r_{ik} in T-UIM. Based on the information source, the node r_{ik} is further subdivided into three distinct sub-nodes r_{ik}^l, r_{ik}^t and r_{ik}^w. The sub-nodes represent the similarity of two tweet sets (TW_i^1, TW_k^2) in content, posting time and location dimension, respectively. The similarity in content dimension is denoted by vector X_{ik}^w. Accordingly, we have vectors X_{ik}^l and X_{ik}^t. A pair of tweet sets is represented as three feature vectors extracted based on three distinct dimensions. The tweet-based user identification problem is converted into the binary classification problem. In other words, the identification of (v_i^1, v_k^2) is changed to the classification of node r_{ik}. We denote the classification result of node r_{ik} by y_{ik}. If $y_{ik} = 1$, the two accounts (u_i^1, u_k^2) belong to the same offline individual; otherwise, these two accounts belong to two distinct person. We can employ the supervised machine learning method to solve the user identification.

4.2 The Framework of T-UIM

The framework of T-UIM is shown in Fig. 2. A pair of user tweet sets (TW_i^1, TW_k^2) is represented as a bag of feature vectors $X_{ik} = \{X_{ik}^l, X_{ik}^t, X_{ik}^w\}$. Assume we have s set of identified users $\{X_{ik}, y_{ik}\}$ for training. Based on the labeled data and the identification model T-UIM, we design a cascaded three-level classifier to identify the user across OSNs.

Feature Extraction Across Networks. In social media, a user often posts hundreds of tweets publicly, containing rich information about: where, when, and what. In the following, we propose to exploit the spatial, temporal and text content information redundancies of two different tweet sets for user identification.

Spatial Features. From the analysis on the tweets posted on different OSNs, we find that (1) users usually post tweets at the same or similar locations, such as home, working places, POIs; (2) these tweets often also mention the same or similar locations. We can utilize the similarity of these locations to identify users. Each location can be specified by geographic latitude and longitude. The similarity of two locations can be represented by their Euclidean distance. For a user on one OSN, we can extract some locations from his public tweets and obtain a set of locations. For a pair of users(v_i^1, v_k^2), we compute the similarities between two sets of locations, and obtain a location similarity matrix $P_{ik}^l = \{p_{mn}\}$, where p_{mn} denotes the Euclidean distance between the m^{th} location of user v_i^1 and the n^{th} location of user v_k^2. We propose to use the Euclidean Distance Distribution of $P_{ik}^l(Euc2D)$ to evaluate the spatial similarity between user v_i^1 and user v_k^2.

Temporal Features. As said in [22], an individual usually post public tweets on different OSNs at similar time slots. Such temporal distribution indicates the user's online activity patterns. For example, some users like to post tweets at night, while other users publish tweets on the way to work. Such users' online patterns are very helpful for user identification. For each user, we can extract the posting time from his public tweets and obtain a sequence of posting time. The similarity of posting time can be represented as the difference in time. Similar to spatial feature, for a pair of users (v_i^1, v_k^2), we can compute the difference between two sequences of time, and get a time similarity matrix $P_{ik}^t = \{p_{mn}\}$, where p_{mn} denotes the difference in the m^{th} posting time of user v_i^1 and the n^{th} time of user v_k^2.Considering the difference in users' online behavior patterns, we compute the time similarity in two different granularities: date and time, and get two corresponding matrices P_{ik}^{t1} and P_{ik}^{t2}. We extract the time difference distribution from P_{ik}^{t1} and P_{ik}^{t2} (DateD and TimeD) to represent the temporal features.

Content Features. We notice that an individual often posts the tweets of similar or same content in different OSNs. This indicates that users usually publish his offline behaviors on multiple different OSNs. These tweets contain many of the same words or synonyms. The similarity of tweet content can be represented by the semantic similarity of content or the number of common words. The tweet content can also help to identify users. Similarly, for a pair of users (v_i^1, v_k^2), we can compute the similarity of two tweets, and get a content similarity matrix $P_{ik}^w = \{p_{mn}\}$. We remove the stop words from tweet, and convert it into a bag-of-words vector. We compute five kinds of similarities: (1) Jaccard coefficient(JacD); (2) the longest common subse-quence(LcsD); (3) the cosine similarity of the two average weight vectors (AwvD); (4) the cosine similarity of the two TFIDF-based weight vectors(TfidfD); (5) the cosine similarity of the two part-of-speech-based weight vectors (PoSD).

Base Classifier Construction. On information dimension $r \in \{l, t, w\}$, we train n base classifier $f_s^r(\cdot)(1 \leq s \leq n)$ with a set of training data $\{X_{ik}^r, y_{ik}\}$. Based on these base classifiers, for a pair of users (v_i^1, v_k^2) and the feature vector X_{ik}^r, we can obtain n confidence score $p_{sr} = f_s^r(X_{ik}^r)(1 \leq s \leq n)$ for user v_i^1 and user v_k^2

belonging to the same user. In practice, the number of unidentified user pairs is much larger than the number of identified users. In these unidentified users, there are also plenty of information redundancies for improving the performance of base classifiers. Following the idea of co-training, we re-train the base classifiers with identified users and unidentified users. After training the base classifiers with identified users, we employ them to identify user pairs on unidentified users. Based on the voting method, we select the unidentified user pairs that more than half of base classifiers agree on the identification result, and put them into training set. We conduct the training process iteratively until convergence. The re-training process is marking out by the (color) dotted line as shown in Fig. 2. The purpose for building n base classifiers is expected to improve identification performance with respect to both accuracy and generalization.

Fusion Classifier Construction. In framework of T-UIM, we design two level fusion classifiers. On information dimension $r \in \{l, t, w\}$, we design the fusion classifier $g_r(\cdot)$ to fusion the classification results of n base classifiers. If a base classifier $f_s^r(\cdot)$ outperforms other base classifiers on dimension r, we take $f_s^r(\cdot)$ as the fusion classifier $g_r(\cdot)$. Otherwise, suppose the classification results of n base classifiers are $\{f_s^r(X_{ik}^r), 1 \leq s \leq n\}$. We train the classifier $g_r(\cdot)$ with a set of data $\{\{f_s^r(X_{ik}^r), 1 \leq s \leq n\}, y_{ik}\}$. Then we use the 2^{nd} level fusion classifier $g'(\cdot)$ to fusion the classification results of $g_l(\cdot)$, $g_t(\cdot)$, $g_w(\cdot)$. The purpose for designing the fusion classifiers is expected to overcome the defect in single base classifier or single information dimension, and generalize the tweet-based user identification.

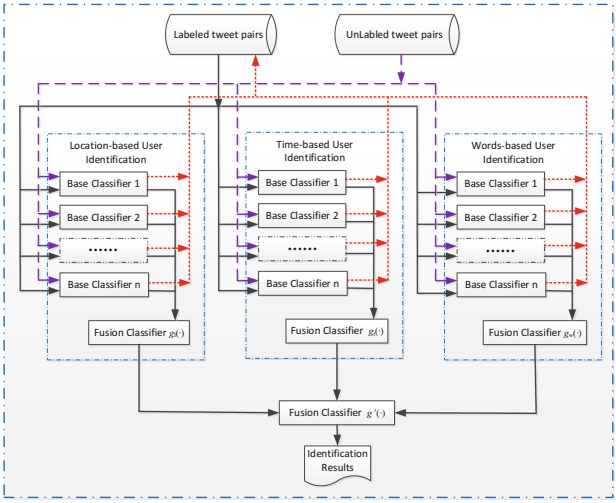


Fig. 2. Framework of T-UIM

5 Experiments and Analysis

Some OSN sites provide the cross-site linking function, such as Foursquare, Google+. Take Foursquare for instance. A user is allowed to make his Facebook and/or Twitter accounts public on his profile page. When a Foursquare user links his profile to his accounts of Facebook and/or Twitter, he should authorize Foursquare to access his Facebook and/or Twitter account. Only after Foursquare verifies the ownership, the user could formally link his public profile to Facebook and/or Twitter account. It is credible to use this method to obtain the ground truth data of a user on different OSN sites. Based on the cross-site function, we could obtain the ground truth data from Facebook, Twitter and Foursquare sites.

Based on the obtain data, we can construct three datasets that only contain positive instances. Three datasets are named as FB-FS, FS-TW and FB-TW, respectively. FS, FB and TW are abbreviations of Foursquare, Facebook and Twitter, respectively. In order to improve the performance of classifiers, we add as many negative instances as positive instances to these three datasets.

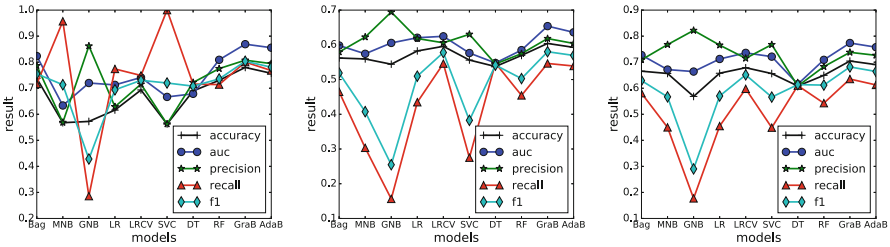
We select ten classifiers including Bagging(Bag), Multinomial Nave Bayes(MNB), Gaussian Nave Bayes(GNB), Logistic Regression(LR), Logistic Regression with builtin Cross-Validation(LRCV), Support Vector Machine (SVM), Decision Tree(DT), Random Forest(RF), GraBoosting(GraB), AdaBoost (AdaB) as base classifiers. The ten base classifiers could be implemented by scikit-learn[1]. All parameters of these classifiers are default values. We perform the 10-fold cross-validation in our experiments. For each dataset, we perform 10 runs, and then report the average of results.

5.1 Comparison and Analysis on Base Classifiers

We first use 10 base classifiers to identify users on datasets FB-TW, FB-FS, and TW-FS. On each dataset, we conduct our experiments on spatial dimension, temporal dimension and content dimension, respectively. To evaluate the performance of 10 base classifiers, we introduce a set of metrics commonly used in machine learning field: Accuracy, Precision, Recall, F1, and AUC.

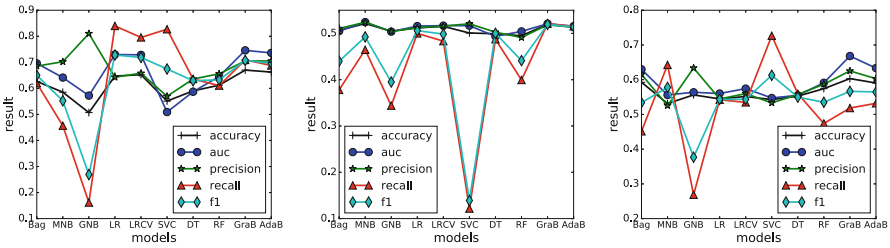
Performance Analysis on Content Dimension. Figure 3 shows the metrics of 10 base classifiers on content dimension. We observe that (1) 10 base classifiers perform best on FB-TW, second on TW-FS, third on FB-FS. Users often post the same activities on Facebook and Twitter simultaneously. Seen from the content, these tweets are more similar. Meanwhile, some users usually recommend delicacies or restaurants to their friends on Twitter while they mark these delicacies or restaurants on Foursquare, but these delicacies or restaurants are often popular and also marked by other users. This causes a little confusion on identifying users. However, due to the different function between Facebook and Foursquare, users rarely simultaneously share the same activities on two sites.

[1] http://scikit-learn.org/stable/.



(a) FB-TW (b) FB-FS (c) TW-FS

Fig. 3. Metric comparison of 10 base classifiers on content dimension.



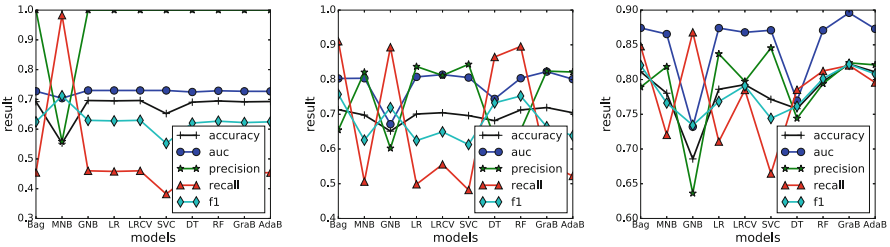
(a) FB-TW (b) FB-FS (c) TW-FS

Fig. 4. Metric comparison of 10 base classifiers on time dimension.

(2) No one base classifier significantly outperforms the other 9 classifiers on three datasets, but GNB significantly do worst on three datasets. For the other 9 classifiers, excluding GNB, each in his own way makes a contribution to identify user. Compared with other classifiers, GraB performs better on three datasets, mainly because GraB is a boosting classifier. (3) All F1 values of 10 classifiers are less than 0.8.

Performance Analysis on Temporal Dimension. Figure 4 shows the metrics of 10 base classifiers on posting time dimension. It illustrates that (1) all base classifiers also perform best on FB-TW, second on FS-TW, and third on FB-FS. The reason for results is same as the reason on content dimension. (2) Compared with results on content dimension, the classifiers perform badly on posting time dimension. At the same date or time, there are many users posting their activities on OSNs. This will lower the identification ability of posting time. (3) The performance of each base classifier on posting time is also not good enough. GraB is with comparatively better results.

Performance Analysis on Location Dimension. Figure 5 shows the metrics of 10 base classifiers on location dimension. We find that (1) 10 base classifiers perform better on location dimension than on other two dimensions. The AUC values are significantly better on location than on other two dimensions. This indicates the location attribute has better identification ability. (2) With the



(a) FB-TW (b) FB-FS (c) TW-FS

Fig. 5. Metric comparison of 10 base classifiers on location dimension.

different results of other two dimensions, the base classifiers perform better on FB-FS and TW-FS than on FB-TW. Because the Foursquare is a location-based OSN, its users provide more locations than other two sites' users. (3) No one classifier outperforms the other 9 classifiers on three datasets significantly. Two boosting classifiers, GraB and AdaB, especially perform worse on FB-FS. This indicates that each classifier has its own merits.

5.2 Analysis on Identification Results with Re-Training

We illustrate the evaluation results of $g'(\cdot)$ after re-training in Table 1. $g'(\cdot) + RT$ represents the evaluation results of $g'(\cdot)$ after re-training. We find that the evaluation results of $g'(\cdot)$ after re-training are slightly better. This indicates that the re-training process is helpful for improving identification results, but its effect is not significant. One reason may be that the features come from the same view of training data. We will conduct our research on re-training the base classifiers across information dimensions in the future work.

Table 1. Evaluation results of $g'(\cdot)$ with re-training process

Dataset	Method	Acc.	Pre.	Rec.	F1
FB-TW	$g'(\cdot)$+RT	**0.8538**	**0.8767**	**0.8591**	**0.8678**
	$g'(\cdot)$	0.8463	0.8661	0.8574	0.8617
FB-FS	$g'(\cdot)$+RT	**0.8597**	**0.8466**	0.8808	**0.8624**
	$g'(\cdot)$	0.8010	0.7455	**0.9292**	0.8256
FS-TW	$g'(\cdot)$+RT	**0.8946**	0.8719	**0.9258**	**0.8979**
	$g'(\cdot)$	0.8663	**0.8787**	0.8485	0.8634

5.3 Comparison with Existing Works

To study the effectiveness of our method, we compare T-UIM with two existing works and their combination: MNA [22], CRMP [24] and MNA+CRMP. For each

Table 2. Evaluation results comparison on T-UIM and the existing works

Dataset	Method	Acc.	Pre.	Rec.	F1
FB-TW	MNA	0.7980	0.8277	0.8062	0.8168
	CRMP	0.6228	0.6820	0.6082	0.6430
	MNA+CRMP	0.8055	0.8323	0.8163	0.8242
	T-UIM	**0.8538**	**0.8767**	**0.8591**	**0.8678**
FB-FS	MNA	0.7418	0.7421	0.7415	0.7415
	CRMP	0.6997	0.7002	0.7665	0.7046
	MNA+CRMP	0.7530	0.7507	0.7583	0.7544
	T-UIM	**0.8597**	**0.8466**	**0.8808**	**0.8624**
FS-TW	MNA	0.8210	0.8202	0.8228	0.8213
	CRMP	0.7453	0.7677	0.7048	0.7345
	MNA+CRMP	0.8484	0.8449	0.8540	0.8493
	T-UIM	**0.8946**	**0.8719**	**0.9258**	**0.8979**

classifier and dataset, we also perform the 10-fold cross-validation. The results are illustrated in Table 2. We find that our method is capable to achieve the best accuracy, recall, precision and F1 on three datasets. This indicates that these suitable features we selected are capable to identify user across OSNs effectively. Besides, it is also shown that MNA and CRMP achieve good performance on three datasets. The combination of MNA and CRMP outperforms better than MNA and CRMP. Three baseline methods perform worse on FB-FS than on other two datasets.

6 Conclusion

In this study we have formalized and studied the problem of user identification across OSNs. As a key and inseparable part of OSN, the tweets posted across OSNs by the same individual usually contain rich information redundancies. This makes the tweet-based user identification possible. Therefore, we proposed a cascaded three-level machine learning-based user identification solution. We developed several algorithms to measure the similarity of tweets on spatial dimension, temporal dimensions and content dimensions. Finally, we verified our solution on three ground truth networks. The results show that our solution can provide excellent performance. Our algorithm could be applied jointly with other profile-based or friendship-based algorithms. The integration of these algorithms is helpful for more accurate identification results.

References

1. Global Social Media Ranking (2017). https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

2. Zhou, X.P., Liang, X., Zhang, H.Y., et al.: Cross-platform identification of anonymous identical users in multiple social media networks. IEEE Trans. Knowl. Data Eng. **28**(2), 411–424 (2016)
3. Liu, J., Zhang, F., Song, X.Y., et al.: What's in a name?: an unsupervised approach to link users across communities. In: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, pp. 495–504 (2013)
4. Zafarani, R., Liu, H.: Connecting users across social media sites: a behavioral-modeling approach. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 41–49 (2013)
5. Iofciu, T., Fankhauser, P., Abel, F., et al.: Identifying users across social tagging systems. In: Proceedings of 5th International AAAI Conference on Weblogs and Social Media, pp. 522–525 (2011)
6. Motoyama, M., Varghese, G.: I seek you: searching and matching individuals in social networks. In: Proceedings of 7th International Workshop on Web Information and Data Management, pp. 67–75 (2009)
7. Abel, F., Herder, E., Houben, G.-J., et al.: Cross-system user modeling and personalization on the social web. User Model. User Adapt. Interact. **23**(2), 169–209 (2013)
8. Raad, E., Dipanda, A., Chbeir, R.: User profile matching in social networks. In: Proceedings of 16th International Conference on Network-Based Information Systems, pp. 297–304 (2010)
9. Vosecky, J., Hong, D., Shen, V.Y.: User identification across multiple social networks. In: Proceedings of 1st International Conferences on Networked Digital Technologies, pp. 360–365 (2009)
10. Jain, P., Kumaraguru, P., Joshi, A.: @ i seek 'fb. me': identifying users across multiple online social networks. In: Proceedings of the 22nd International Conference on World Wide Web Companion, pp. 1259–1268 (2013)
11. Vosecky, J., Hong, D., Shen, V.Y.: User identification across social networks using the web profile and friend network. Int. J. Web Appl. **2**(1), 23–34 (2010)
12. Buccafurri, F., Lax, G., Nocera, A., Ursino, D.: Discovering links among social networks. In: Flach, P.A., Bie, T., Cristianini, N. (eds.) ECML PKDD 2012. LNCS (LNAI), vol. 7524, pp. 467–482. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33486-3_30
13. Tan, S., Guan, Z.Y., Cai, D., et al.: Mapping users across networks by manifold alignment on hypergraph. In: Proceedings of 28th AAAI Conference on Artificial Intelligence, pp. 159–165 (2014)
14. You, G.-W., Hwang, S.-W., Nie, Z.Q., et al.: SocialSearch: enhancing entity search with social network matching. In: Proceedings of the 14th International Conference on Extending Database Technology, pp. 515–519 (2011)
15. Goga, O.: Matching user accounts across online social networks: methods and applications. Ph.D. Dissertation, Universite Pierre etmarie curie – Pairs 6, Franch (2014)
16. Vesdapunt, N., Hector, G.-M.: Identifying users in social networks with limited information. In: Proceedings of the IEEE 31st International Conference on Data Engineering, pp. 627–638 (2015)
17. Huang, S.R., Zhang, J., Lu, S.Y., et al.: Social friend recommendation based on network correlation and feature co-clustering. In: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, pp. 315–322 (2015)
18. Zafarani, R., Tang, L., Liu, H.: User identification across social media. ACM Trans. Knowl. Discov. Data **10**(2), 1–30 (2015)
19. Shmatikov, V., Narayanan, A.: De-anonymizing social networks. In: Proceedings of IEEE Symposium on Security and Privacy, pp. 173–187 (2009)

20. Bartunov, S., Korshunov, A., Park, S.-T., et al.: Joint link-attribute user identity resolution in online social networks. In: Proceedings of 6th SNA-KDD Workshop (2012)
21. Korula, N., Lattanzi, S.: An efficient reconciliation algorithm for social networks. Proc. VLDB Endow. **7**(5), 377–388 (2013)
22. Kong, X.N., Zhang, J.W., Yu, P.-S.: Inferring anchor links across multiple heterogeneous social networks. In: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, pp. 179–188 (2013)
23. Goga, O., Lei, H., Hari, S., et al.: Exploiting innocuous activity for correlating users across sites. In Proceedings of the 22nd International Conference on World Wide Web, pp. 447–458 (2013)
24. Sajadmanesh, S., Rabiee, H.R., Khodadadi, A.: Predicting anchor links between heterogeneous social networks. In: Proceedings of 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 158–163 (2016)
25. Zhang, J.W., Kong, X.N., Yu, P.-S.: Predicting social links for new users across aligned heterogeneous social networks. In: Proceedings of IEEE 13th International Conference on Data Mining, pp. 1289–1294 (2013)
26. Jain, P., Kumaraguru, P.: Finding nemo: searching and resolving identities of users across online social networks. arXiv preprint 2012. arxiv:1212.6147
27. Perito, D., Castelluccia, C., Kaafar, M., et al.: How unique and traceable are usernames? In: Proceedings of 11th International Conference on Privacy Enhancing Technologies, pp. 1–17 (2011)
28. Jiang, X., Wei, S.K., Zhao, R.Z., et al.: Camera fingerprint: a new perspective for identifying user's identity. arXiv preprint arxiv: 1610.07728 (2016)
29. Liu, D., Wu, Q.Y., Han, W.H.: User identification across multiple websites based on usern-ame features. Chin. J. Comput. **38**(10), 2028–2040 (2015)
30. Bennacer, N., Nana Jipmo, C., Penta, A., Quercini, G.: Matching user profiles across social networks. In: Jarke, M., Mylopoulos, J., Quix, C., Rolland, C., Manolopoulos, Y., Mouratidis, H., Horkoff, J. (eds.) CAiSE 2014. LNCS, vol. 8484, pp. 424–438. Springer, Cham (2014). doi:10.1007/978-3-319-07881-6_29
31. Malhotra, A., Totti, L., Meira, W., et al.: Studying user footprints in different online social networks. In: Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining, pp. 1065–1070 (2012)
32. Almishari, M., Tsudik, G.: Exploring linkability of user reviews. In: Foresti, S., Yung, M., Martinelli, F. (eds.) ESORICS 2012. LNCS, vol. 7459, pp. 307–324. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33167-1_18