ChinaRxiv

AI translation · View original & related papers at
chinarxiv.org/items/chinaxiv-201812.00123

Cross-Social Network User Identity Linkage Based on User Relationships (Postprint)

Authors: Liu Qifei, Du Yanhui, Lu Tianliang

Date: 2018-12-13T00:00:00+00:00

Abstract

To identify accounts belonging to the same natural person across different social network platforms, a cross-social-network user identity association method based on user relationships is proposed. First, a user relationship extraction module based on network representation learning is designed, which transforms large-scale user relationships into a low-dimensional vector space for representation. Then, traditional network representation learning algorithms are improved for heterogeneous information networks, and the CSN_LINE algorithm is proposed to achieve network representation that integrates cross-social-network prior association relationships. Finally, a user identity association model based on multi-layer perceptron is constructed. Experimental results show that, compared with current state-of-the-art methods, the improvements in both the comprehensive metric F1-score and accuracy exceed 12%, demonstrating the rationality and effectiveness of the proposed method.

Full Text

Preamble

Vol. 37 No. 2
Application Research of Computers
ChinaXiv Partner Journal

User Identity Linkage Across Social Networks Based on User Relations

Liu Qifei , Du Yanhui , , Lu Tianliang , †

(a. Information Technology & Network Security Institute; b. CIC of Security & Law for Cyberspace, People' s Public Security University of China, Beijing 100038, China)

Abstract: To identify accounts belonging to the same natural person across different social networking platforms, this paper proposes a cross-social network user identity linkage method based on user relations. First, we design a user relation extraction module based on network representation learning, which embeds large-scale user relations into a low-dimensional vector space. Then, we improve the traditional network representation learning algorithm for heterogeneous information networks and propose the CSN_LINE algorithm to achieve network representation that fuses cross-network anchor links. Finally, we construct a user identity linkage model based on a multilayer perceptron. Experimental results demonstrate that compared with current state-of-the-art methods, the proposed method achieves improvements of over 12% in both the comprehensive F1-score and accuracy, proving its rationality and effectiveness.

Keywords: user relations; cross-social networks; user identity linkage; network representation learning; multilayer perceptron

0 Introduction

Recent research by GlobalWebIndex indicates that globally, 98% of internet users utilize at least one social network, with an average of 7.6 online accounts per user. As users commonly maintain identities across multiple social networking platforms, their information becomes fragmented across these disparate services. Breaking down these "information silos" to achieve multi-source heterogeneous data fusion necessitates cross-social network user identity linkage, which provides richer data support for complex social network analysis tasks. For instance, it enables more comprehensive user profiling, helps recommendation systems deliver more precise personalized services, solves the "cold start" problem, and supports cybersecurity efforts such as identifying fake or illegal accounts. Consequently, cross-social network user identity linkage represents a highly significant research area with broad application value.

Numerous studies have focused on user identity linkage through user attribute and behavior information, achieving notable results. However, growing privacy concerns make such information increasingly difficult to obtain and verify, posing significant challenges. To overcome these difficulties, leveraging user relations offers a promising alternative. User relationship topology is inherently anonymous, and ordinary users rarely fabricate meaningless connections in their personal accounts, making user relations a more authentic reflection of reality. These relationships embody user emotions and interests while mirroring real-world social connections. Identity linkage based on social network relations can compensate for limitations of attribute- and behavior-based approaches, enhancing robustness and generalization capability.

Nevertheless, research on user identity linkage based on user relations faces several challenges: (a) quantitatively representing user relations and capturing topological features is difficult; (b) with massive user bases on social platforms,

developing efficient multi-account linkage algorithms for large-scale complex networks remains a challenge; and (c) due to the scale-free and small-world properties of social networks, user relationship topologies exhibit high homogeneity, making precise identity linkage algorithmically challenging.

To address these issues, this paper proposes a cross-social network user identity linkage method based on user relations. The main contributions are: (a) utilizing network representation learning to design a user relation feature extraction method that transforms user relation features into low-dimensional vector representations; (b) improving traditional network representation learning algorithms for heterogeneous information networks in cross-social network scenarios, resulting in the CSN_LINE algorithm; (c) constructing a user identity linkage model based on multilayer perceptrons; and (d) collecting large-scale user relation data from Sina Weibo and Douban to train and validate the proposed method.

---

3 Cross-Social Network User Identity Linkage Method

Researchers typically design cross-social network user identity linkage methods across three dimensions: user attributes, user behavior, and user relations. User attribute features primarily include username, personal description, gender, occupation, and profile picture [1–4], while user behavior features mainly encompass writing style [5–7] and user location trajectories [8,9]. User relations have also been extensively studied.

Liu et al. [10] proposed the HYDRA multi-account linkage method by modeling user behavior through long-term analysis and short-term multi-angle information matching, while simultaneously leveraging the structural isomorphism of user ego-networks. Tan et al. [11] represented network relations as matrices using hypergraphs and reduced matrix dimensionality to decrease computational complexity. While user relations can generally be represented via adjacency matrices, these matrices become extremely sparse in large-scale networks. Man et al. [12] introduced the PALE method, which employs network representation learning to map user nodes into low-dimensional vector spaces for identity linkage. Feng et al. [13] designed two novel methods for measuring similarity between users across different social networks. Zhang et al. [14] proposed COSNET, which integrates local and global matching information from social network topology graphs and utilizes energy models to solve multi-account linkage problems. Zhou et al. [15] presented the FRUI method, which significantly reduces time complexity by fully leveraging pre-linked cross-social network user pairs. Wang et al. [16] employed crowdsourcing to increase training sample size and utilized full-view features to measure user similarity, proposing a cross-social network user identification method combining comprehensive features with crowdsourcing.

---

2 Problem Analysis

The objective of user identity linkage based on user relations is to identify accounts belonging to the same natural person across different social networking platforms. We formalize this problem as follows.

Consider two distinct social networking platforms, denoted as G and G , where G = (V , E) and G = (V , E). V and V represent the user sets in platforms A and B, respectively, while E and E represent the user relation sets within each platform. The cross-network anchor links are defined as M = {(v, u) | v V , u V }, where M contains user pairs from different platforms belonging to the same natural person.

As illustrated in [Figure 1: see original paper], within platforms A and B, users have intra-platform connections (solid lines in Figure 1), such as follow or friend relationships. Additionally, some prior anchor links exist between platforms (three dashed lines at the top of Figure 1), representing pre-identified user pairs belonging to the same person. Leveraging both intra-platform relations and cross-platform anchor links, user identity linkage based on user relations aims to discover additional unknown cross-network connections (two dashed lines at the bottom of Figure 1).

The proposed method comprises two key components: a user relation feature extraction module and a multilayer perceptron-based user identity linkage model. The overall workflow is shown in [Figure 2: see original paper]. The feature extraction module vectorizes user node topology information, representing topological features through low-dimensional dense vectors. The MLP-based linkage model employs a multilayer perceptron to train a binary classifier that determines whether user pairs from different social networks correspond to the same person.

4.1 Feature Extraction Based on Network Representation Learning

User relations constitute a fundamental characteristic of social networking platforms. Complex inter-user connections transform individual users into network communities, forming social networks. Across different platforms, user relations carry different semantics, primarily categorized as follow relationships (directed graphs) or friend relationships (undirected graphs). For instance, Sina Weibo represents a typical directed social network where connections are one-way follow relationships, while Facebook exemplifies an undirected social network where connections require mutual confirmation.

To achieve user identity linkage based on user relations, we must transform these relations into features readable by downstream linkage models. Network representation learning methods represent nodes as low-dimensional dense vectors,

which can serve as inputs for subsequent analysis models, as shown in [Figure 3: see original paper].

This paper employs the popular LINE algorithm [17] to generate low-dimensional dense vector representations of user nodes, as it accommodates both directed and undirected graphs. However, our research object is not a traditional network structure but rather a heterogeneous information network spanning multiple social platforms. As illustrated in the problem analysis (Figure 1), the network originates from two distinct platforms, resulting in two node types and two relationship types. Traditional network representation learning methods focus on single-platform intra-network structures and cannot capture cross-social network relationships. To fully leverage cross-network anchor links and enhance applicability to heterogeneous information networks, we propose the CSN_LINE algorithm.

4.2 Traditional LINE Algorithm

The traditional LINE algorithm defines first-order and second-order proximity. First-order proximity represents the direct intimacy between two nodes. For connected nodes v and v through edge (i, j) with weight w , w reflects their first-order proximity; if no edge connects v and v , their first-order proximity is zero.

By modeling the actual and empirical probability distributions between nodes v and v , the algorithm preserves first-order proximity characteristics by minimizing the difference between these distributions, measured via KL divergence. The first-order proximity objective function is given by equation (1), where P (v , v) represents the actual probability modeling of first-order proximity.

Second-order proximity captures indirect intimacy through the number of shared neighbor nodes, assuming that nodes with many common neighbors are more similar. If no node simultaneously connects to v and v , their second-order proximity is zero. Similar to first-order proximity, the algorithm minimizes the KL divergence between distributions, with the second-order proximity objective function given by equation (2).

By minimizing the objective functions in equations (1) and (2), the algorithm generates vector representations that preserve both first-order and second-order proximity characteristics between nodes.

4.3 CSN_LINE Algorithm

To address cross-social network user identity linkage scenarios and fully exploit inter-platform relationships for heterogeneous information network representa-

tion learning, we propose CSN_LINE, which incorporates prior anchor links into first-order proximity.

For a cross-network anchor link (v, u) where nodes v and u originate from different platforms, the actual probability distribution is modeled in equation (3). Since anchor links represent known same-person account pairs, their importance should substantially exceed that of intra-platform edges. Therefore, we introduce an adjustment parameter in the empirical probability distribution formula, as shown in equation (4), where W represents the sum of all edge weights in the network (or simply the edge count for unweighted networks). If two nodes are connected, w = 1; otherwise, w = 0.

Consequently, the first-order proximity objective function incorporating prior anchor links is given by equation (5). Minimizing this objective function ensures node vector representations capture cross-network anchor relationship features.

The proposed CSN_LINE algorithm optimizes the objective functions in equations (1), (2), and (5) using stochastic gradient descent to learn node vector representations, ultimately producing low-dimensional vectors for all nodes across both social networks. These vectors reflect not only intra-platform connections but also inter-platform anchor relationships, serving as comprehensive user relation features.

---

5 MLP-Based User Identity Linkage Model

Determining whether two users from different social networks belong to the same natural person can be formulated as a binary classification problem. The input consists of feature vectors from two users across different platforms, with output classification results of 1 or -1, where 1 indicates the same person and -1 indicates different individuals.

In the cross-social network user identity linkage model, we employ a multilayer perceptron (MLP) as the classifier. For binary classification, the output layer contains two neurons. The user node vectors from the first and second social networks are concatenated into a single long vector that serves as MLP input, with the number of input neurons equal to the concatenated vector' s dimension. As shown in [Figure 4: see original paper], the hidden layer count and neuron quantities are illustrative rather than prescriptive.

We train the MLP network using concatenated vectors from known same-person account pairs as positive samples (labeled 1) and concatenated vectors from randomly selected non-matching account pairs as negative samples (labeled -1). The trained classifier functions as the user identity linkage model.

6.1 Dataset Description

Collecting datasets for cross-platform user identity linkage is challenging due to privacy protections, as it is nearly impossible to verify same-person accounts through private data like phone numbers or email addresses. Veiga et al. [18] proposed a data collection method that leverages user-generated content to find cross-platform clues, such as when users post links to their profiles on other platforms. This approach was successfully applied to Twitter, Instagram, and Foursquare.

Following Veiga' s methodology, we collected large-scale user relation data from Douban and Sina Weibo. The dataset statistics are presented in . The dataset contains 14,457 cross-network anchor user pairs, meaning 14,457 Douban accounts and 14,457 Sina Weibo accounts belong to 14,457 distinct natural persons.

For binary classifier training, positive samples consist of concatenated network representation vectors from the 14,457 anchor pairs (labeled 1). We randomly select 14,457 non-matching account pairs, concatenate their network representation vectors as negative samples (labeled -1), resulting in 28,914 total training samples.

6.2 Evaluation Metrics

We employ standard evaluation metrics including precision (P), recall (R), F1-score (F1), and accuracy (Acc), calculated using equations (9)-(12). Here, tp represents true positives, fp false positives, tn true negatives, and fn false negatives.

6.3 Comparison of Common Network Representation Learning Algorithms

Selecting an appropriate network representation learning method is crucial for leveraging user relation features in cross-social network identity linkage. We implemented several popular algorithms: Deepwalk, LINE, and Node2vec, including three LINE variants (first-order only, second-order only, and both). All node representation vectors were set to 50 dimensions, with Node2vec random walk parameters configured as p=0.25, q=0.25.

To evaluate linkage performance under different representation methods, we configured the MLP with 2 hidden layers of 200 neurons each, implemented using the scikit-learn machine learning library. During training, 70% of the dataset served as the training set and 30% as the test set. The results are shown in .

LINE algorithms fundamentally differ from Deepwalk and Node2vec: LINE optimizes proximity-based objective functions to generate node vectors, whereas Deepwalk and Node2vec employ random walks to generate node sequences followed by word2vec-style neural network training. The experimental results demonstrate that LINE using both first-order and second-order proximity achieves the best performance for cross-social network user identity linkage.

6.4 Effectiveness Validation of CSN_LINE Algorithm

We improved the traditional LINE algorithm by incorporating prior anchor links into first-order proximity, adding a third objective function to enable heterogeneous information network representation learning for cross-social networks. To validate CSN_LINE' s effectiveness, we conducted comparative experiments with two adjustable parameters.

The first parameter is the adjustment coefficient in equation (5). We tested values of 0, 3, 5, 7, and 9, where =0 represents standard LINE without anchor link incorporation. Using all 10,120 anchor links from the training set, the comparative results are presented in .

Analysis reveals that values of 5, 7, and 9 maintain relatively high linkage performance, =3 yields moderate results, and =0 performs worst. Compared with traditional LINE (=0), CSN_LINE (=3,5,7,9) demonstrates superior performance, validating the effectiveness of our improved algorithm.

To further demonstrate the contribution of anchor link incorporation, we conducted additional experiments with =5 while varying the percentage of anchor links used (25%, 50%, 75%, 100%). The results are shown in [Figure 5: see original paper], indicating that higher numbers of anchor links lead to better F1-scores and accuracy. This confirms that anchor links contribute positively to user identity linkage performance, and network representation learning incorporating prior anchor links enhances linkage effectiveness.

6.5 Method Comparison

We compare our proposed method with two representative approaches to demonstrate its superiority.

The first is an unsupervised method based on counting common neighbor nodes. The core idea assumes that if two users from different platforms share many cross-platform linked neighbor pairs, they likely belong to the same person. Zhong et al. [19] proposed CoLink, which counts cross-platform linked neighbor pairs for two users and compares this count against a threshold to determine identity linkage. Similar strategies were employed by Sun et al. [4] and Qi [20].

The second approach uses network representation learning for low-dimensional vector representation followed by supervised machine learning for model training. Man et al. [12] proposed PALE, which uses LINE' s first-order proximity for node representation and a single-hidden-layer MLP for linkage.

Comparative experiments on our dataset, shown in [Figure 6: see original paper], demonstrate that our proposed method outperforms both CoLink and PALE across all four evaluation metrics. CoLink' s limitations stem from its overly simplistic single-dimensional features and threshold-based decision making, which is insensitive to local neighbor differences and yields low precision with high recall. PALE' s inferior performance arises because traditional network representation methods struggle with heterogeneous information networks, producing node vectors that cannot capture cross-network relationships as effectively as our approach.

---

7 Conclusion

Fusing multi-source heterogeneous data from social networking platforms requires linking accounts belonging to the same natural person across services. This paper proposes a cross-social network user identity linkage method based on user relations, employing network representation learning to encode user relation features into low-dimensional dense vectors. We improve the traditional LINE algorithm for heterogeneous information networks in cross-social network scenarios by incorporating prior anchor links, and construct a user identity linkage model using multilayer perceptrons. Training and testing on large-scale real-world user relation data fully demonstrate the method' s effectiveness and its applicability to social network user data fusion tasks.

Future work will focus on collecting more comprehensive user data to design integrated multi-dimensional feature models combining user attributes, user behavior, and user relations, thereby further improving identity linkage performance.

References

[1] Liu Dong, Wu Quanyuan, Han Weihong, et al. User Identification across multiple websites based on username features [J]. Chinese Journal of Computers, 2015, 38(10): 2028-2040.

[2] Zafarani R, Tang L, Liu H. User identification across social media [J]. ACM Trans on Knowledge Discovery from Data, 2015, 10(2).

[3] Wu Zheng, Yu Hongtao, Liu Shuxin, et al. User identification across multiple social networks based on information entropy [J]. Journal of Computer Applications, 2017, 37(8): 2374-2380.

[4] Sun Song, Li Qiudan, Yan Peng, et al. Mapping users across social media platforms by integrating text and structure information [C]//Proc of IEEE International Conference on Intelligence and Security Informatics. 2017: 113-118.

[5] Vosoughi S, Zhou H, Roy D. Digital Stylometry: Linking Profiles Across Social Networks [C]//Proc of International Conference on Social Informatics. 2015: 164-177.

[6] Sha Ying, Liang Qi, Zheng Kaijiang. Matching user accounts across social networks based on users message [J]. Procedia Computer Science, 2016, 80: 2423-2427.

[7] Li Yongjun, Zhang Zhen, Peng You, et al. Matching user accounts based on user generated content across social networks [J]. Future Generation Computer Systems, 2018, 83: 104-115.

[8] Kong Xiangnan, Zhang Jiawei, Yu PhilipS. Inferring anchor links across multiple heterogeneous social networks [C]//Proc of ACM International Conference on Conference on Information and Knowledge Management. New York: ACM Press, 2013: 179-188.

[9] Zhang Jiawei, Kong Xiangnan, Yu P S. Transferring heterogeneous links across location-based social networks [C]//Proc of ACM International Conference on Web Search and Data Mining. New York: ACM Press, 2014: 303-312.

[10] Liu Siyuan, Wang Shuhui, Zhu Feida, et al. HYDRA: large-scale social identity linkage via heterogeneous behavior modeling [C]//Proc of ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2014: 51-62.

[11] Tan Shulong, Guan Ziyu, Cai Deng, et al. Mapping users across networks by Manifold Alignment on Hypergraph [C]//Proc of the 28th AAAI Conference on Artificial Intelligence. 2014: 159-165.

[12] Man Tong, Shen Huawei, Liu Shenghua, et al. Predict anchor links across social networks via an embedding approach [C]//Proc of International Joint Conference on Artificial Intelligence. 2016.

[13] Feng Shuo, Shen Derong, Kou Yue, et al. Anchor link prediction using topological information in social networks [C]//Proc of the 17th International Conference on Web-Age Information Management. [S.l.]: Springer International Publishing. 2016: 338-352.

[14] Zhang Yutao, Tang Jie, Yang Zhilin, et al. COSNET: connecting heterogeneous social networks with local and global consistency [C]//Proc of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2016: 1485-1494.

[15] Zhou Xiaoping, Liang Xun, Zhang Haiyan, et al. Cross-platform identification of anonymous identical users in multiple social media networks [J]. IEEE Trans on Knowledge and Data Engineering, 2016, 28(2): 411-424.

[16] Wang Qian, Shen Derong, Feng Shuo, et al. Identifying users across social networks based on global view features with crowdsourcing [J]. Journal of Software, 2018, 29(3): 811-823.

[17] Tang Jian, Qu Meng, Wang Mingzhe, et al. LINE: large-scale information network embedding [C]//Proc of the 24th International Conference on World Wide Web. New York: ACM Press, 2015.

[18] Veiga M H, Eickhoff C. A Cross-Platform Collection of Social Network Profiles [C]//Proc of International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2016: 665-668.

[19] Zhong Zexuan, CaoYong, Guo Mu, et al. CoLink: An Unsupervised Framework for User Identity Linkage [C]//Proc of AAAI Conference on Artificial Intelligence. 2018.

[20] Qi Linfeng. The identity of the same user with cross-social media based on entity resolution [J]. Library and information service, 2017, 61(6): 107-114.

Note: Figure translations are in progress. See original paper for figures.

Source: ChinaXiv —Machine translation. Verify with original.