Cross-domain entity identity association analysis and prediction based on representation learning

Mingcheng Gao¹,², Ruiheng Wang¹,², Lu Wang¹,², Yang Xin¹,² and Hongliang Zhu¹,²

Abstract
Cross-domain identity association of network entities is a significant research challenge and a vital issue of practical value in relationship discovery and service recommendation between things in the Internet of things, cyberspace resources surveying mapping, threat tracking, and intelligent recommendation. This task usually adds additional difficulty to the research in practical applications due to the need to link across multiple platforms. The existing entity identity association methods in cross-domain networks mainly use the attribute information, generated content, and network structure information of network user entities but do not fully use the inherent strong positioning characteristics of active nodes in the network. In this article, we analyzed the structural characteristics of existing relational networks. We found that the hub node has the role of identity association positioning, and the importance of identity association reflected by different nodes is different. Moreover, we creatively designed a network representation learning method. We proposed a supervised learning identity association model combined with a representation learning method. Experiments on the public data set show that using the identity association method proposed in this article, the ranking accuracy of user entity association similarity is about 30% and 25% higher than the existing two typical methods.

Keywords
Cross-domain network, network entity, identity association, representation learning, hub node, Internet of things

Date received: 21 January 2022; accepted: 12 September 2022

Handling Editor: Peio Lopez Iturri

Introduction
With the gradual expansion of the extension of cyber-space, mobile applications, Internet of things (IoT) applications, big data, and other applications are becoming more and more widely and have attracted more and more attention. A user entity usually has different identities or accounts in the IoT, social networks, e-commerce networks, and other networks, affecting people’s lives and work. For example, intelligent life centered on user entities in the IoT, social activities on Facebook, Twitter, Instagram, YouTube, Weibo, WeChat, TikTok, and other media, also online shopping activities such as Amazon, Tmall, and JD.com.

In practical application, the association analysis of network entities faces the problems of cross-platform and privacy protection, which brings difficulties to the relevant research. The association analysis of
cross-domain network user entities has important practical significance in the fields of network behavior analysis and prediction, relationship discovery and intelligent service recommendation of the IoT, network behavior traceability, cyberspace resources surveying mapping, information dissemination, viral marketing, cross-domain intelligent recommendation, and so on. Identity association is essentially the closest matching problem of network entities in multiple network domains. The final output of the identity association model is the similarity between network entities; the more significant the similarity, the greater the probability that they are the same physical entity (in the social media domain, the physical entity is a natural person). In the field of intelligent recommendation of the IoT, the greater the similarity means that, the higher the similarity of usage habits between users, the greater the significance of service intelligent recommendation.

Current research, such as in the social media domain, mainly focuses on combining supervised and unsupervised methods to realize the identity association of user entities using the attribute information, generated contents, and network structure information of network users. (1) Methods based on user attributes, such as Zafarani and Liu, analyze some user identity links, use naive Bayesian classifier, support vector machine, and so on, and combine the behavior pattern constructed by the psychological analysis method to associate user identity. (2) Based on the user-generated content method, learn and extract the special representation of user identity by extracting user-generated content features, such as the characteristic vocabulary, emotion, expression mode, and track of message content recognition. (3) Previous works and others use information such as network structure features and user attribute features to learn the connectivity of network topology through graph structure, mine the similarity of identity features between and within network domains, and identify the identity of unknown cross-domain network user entities by comparing one by one and iterating many times.

Among the existing identity association methods, supervised and semi-supervised methods can obtain better correlation accuracy than unsupervised methods. In the identity association analysis method of network entities based on network structure, most of the existing studies focus on the user identity link based on user anchor node and predict the associated users in the network through anchor node user detection and network identity link analysis.

Generally speaking, the current association analysis method of network entity identity mainly realizes the identity association of cross-domain network user entities based on representation learning method and end-to-end method based on deep learning. The representation learning method maps the user identity to a unified space based on the deep extraction of user information and the accurate expression of user characteristics; determine whether the user identity is associated through the analysis of association degree. Based on the end-to-end method of deep learning, we use the characteristic information of identity association, such as the generated content of users, as well as attribute information such as nickname, network location, social relationship, and so on, into the deep learning network, directly calculate the association similarity between users and give the probabilistic judgment result of identity association. Researchers have made many efforts to solve the cross-domain user identity association problem. However, there are still difficulties in effectively using the network structure to recognize the relationship between user entities.

According to the above analysis, to meet the requirements of accuracy, the association analysis of cross-domain network user entity identity based on network structure depends not only on the user’s attributes, neighbors, and interaction information with neighbors but also on the structural features of the cross-domain network. In essence, identity association is in opposition to privacy protection. In the field of privacy protection, hub nodes have intense identity exposure in structural features. These hub nodes often have extensive and frequent interpersonal interactions and can form different social groups centered on these nodes. This phenomenon of the hub node has strong recognition in the field of entity identity association prediction. Still, the existing association methods based on network structure representation learning usually ignore this information. This article designs a representation learning method for network structure, which can learn the location characteristics and essential characteristics of nodes, set attention mechanisms to improve the representation ability, and propose an identity association model.

The main contributions of this article are summarized as follows:

1. We analyzed the structural features of the Twitter–Foursquare social network and found that (1) different nodes reflect different positioning functions, among which the nodes with a higher degree have stronger identity association positioning function; (2) different adjacent nodes reflect different importance degrees of identity association, and we creatively designed a network representation learning method, which uses nodes, respectively, location feature and importance feature are represented by two sets of vectors.

2. We propose an identity association model for supervised learning IFN-UIL (importance features of nodes and user identity linkage (UIL));
the model combines the essential features of nodes to improve the accuracy of user identity association analysis in complex networks.

3. Using Twitter and Foursquare data sets, we evaluate the effect of identity association through two methods: P@N and H@N, and compare them with existing identity association methods. The results show that the proposed identity association method has certain advanced nature.

The other parts of this article are organized as follows. The “Related work” section summarizes the related work. The “Problem definition” section formalizes the identity association in cross-domain networks. The “Methods” section analyzes the structural features of the Twitter–Foursquare social network and expounds on the details of the proposed network representation learning method and identity association model. The “Experiments” part is experimental evaluation. The “Conclusion” part is the conclusion of this article.

**Related work**

In recent years, network embedding has aroused extensive research interest. Network embedding aims to learn the low dimensional representation of network nodes and effectively retain the edge information such as network topology and node content. Cross-domain network embedding is a relatively new research problem at present. It has crucial research significance in intelligent recommendation of IoT services, abnormal network behavior analysis, intelligent recommendation of e-commerce network, network traceability, cyberspace resources surveying mapping, and so on.

In social and e-commerce networks, network structure information is a networked representation of social relations between different users, such as friend relationships, family friend relationships, colleague relationships, subordinate relationships, and follow-up relationships. The network structure information is easier to obtain than user attributes information. Existing studies have proposed many methods on network representation learning combined with identity association. Tan et al.\(^1\) integrated two social networks into a complete network, mapped it to a hypergraph, and used multi neighborhood relations to learn more useful potential network features. Zhou et al.,\(^1\) based on the DeepWalk method,\(^2\) encoded the user nodes in the network into vector representation to capture the local and global network structures. These structures can intensely train the user identity association model between different network domains based on semi-supervised methods. Translation-based technology has significant advantages in representing network nodes and edges in complex networks. It can embed the cross-domain network user information and the diverse interaction relationship between users into the low latitude vector space to establish the relationship representation model.\(^3\) Miao et al.\(^2\) and Wang et al.\(^4\) proposed a user identity association method based on representation learning based on the network structure information, and some anchor links are known. The observed anchor links are used to train the mapping function of identity association prediction and find more hidden anchor links between the two social networks.

In IoT, the discovery method of inter-entity relationship based on graph embedding, such as Yao and colleagues,\(^5\) maps the context information of nodes to a separate graph to capture the association relationship between different entities. Yao et al.\(^6\) expressed the complex relationship between entities centered on events by constructing hypergraphs, in which vertices represent entities in the microdomain of the IoT and hyperedges define heterogeneous relationships among entities.

However, the current methods encounter bottlenecks in anchor user determination and identity association combination discovery.\(^7\) They lack sufficient association degree support conditions between nodes, resulting in the low accuracy of entity identity association prediction in complex networks. These methods focus more on preserving the generated embedding vector’s structure information and node proximity information. This information is used for the subsequent identity association mining tasks in the network. Whether the DeepWalk method or Node2Vec method is used in the node embedding, due to the inherent characteristics of these methods, they ignore the strong positioning characteristics of the central node in the network.

**Problem definition**

We first define and introduce the concepts and symbols used in this article.

**Network**

A network is represented as a graph \(G = \{V, E\}\), where \(|V|\) and \(|E|\) represent the number of vertices and edges, respectively. \(V\) denotes the set of nodes, \(v_i \in V\) represents a network entity, \(E = V \times V\) denotes the set of links, which represents the relationship between entities. Network entities include user entities, legal entities, equipment entities, and application entities.

**UIL**

Given any two social networks \(G_i\) and \(G_j\), the goal of UIL is to predict that any pair of identities \(v_i\) and \(v_j\) are
chosen from $G_i$ and $G_j$, respectively, belong to the same real natural person, legal person, equipment, or application (i.e. $u_i = u_j$). Mathematically, UIL is to find a binary function $f$ such that

$$f(v_i, v_j) = \begin{cases} 1, & v_i, v_j \text{ as the same natural person, legal person, equipment or application} \\ 0, & \text{otherwise} \end{cases}$$

The binary function $f(x)$ that is defined decides perfectly if a set of entities on various network domain correspond to the same real person, legal person, equipment, or application.

**Methods**

Taking social networks as an example, this article discusses the entity identity association combined with hub nodes degree information in the Twitter domain and Foursquare domain.

**Data set analysis**

The purpose of data set analysis is to find the defects of current research methods, which is the basis of this research idea. We use the Twitter–Foursquare data set\(^3\) in the social network to analyze the association of network user entity identity. The data set includes the number of user nodes, the number of user relationships, and the number of user combinations belonging to the same identity between the Twitter and Foursquare domains, as shown in Table 1. We use the overlap similarity calculation model. First, the degree distribution characteristics of user nodes in the Twitter–Foursquare data set are analyzed. The results show that the user nodes with a high degree have a strong role of identity association and positioning in cross-domain networks. Second, the similarity between different adjacent nodes of user node $V$ in the Twitter–Foursquare data set is analyzed. The results show that the direct neighbor nodes have higher support for node $V$ identity association.

| Social networks | Number of users | Relationship quantity | Number of identity association combinations |
|------------------|-----------------|----------------------|--------------------------------------------|
| Twitter          | 5220            | 164,919              | 1609                                       |
| Foursquare       | 5313            | 76,972               |                                             |

Table 1. Experiment and evaluation training data.

In Twitter and Foursquare social networks, the user nodes in ground truth are sorted in degree characteristics from large to small and are divided into 10 groups of node-sets on average. Use the following formula to calculate the same natural person coverage of Twitter and Foursquare node-sets

$$s(V^T_i, V^F_i) = \frac{|F(V^T_i) \cap F(V^F_i)|}{\min(|F(V^T_i)|, |F(V^F_i)|)} \quad (2)$$

where $i \in [1, 10]$ represents the number of the combination, $V^T_i$ represents the set of nodes in group $i$ of Twitter, $V^F_i$ represents the set of nodes in group $i$ of Foursquare, and $F(V)$ represents the set of natural persons corresponding to the node-set $V$, the associated user combination or set. $|F(V^T_i) \cap F(V^F_i)|$ refers to the number of natural persons who appear in Twitter and Foursquare network domains at the same time in the node-set of group $i$, the minimum number of natural persons appearing in the Twitter network domain or Foursquare network domain in the node-set of group $i$, and the maximum value of similarity is 1.

In different degree ranking sets, As shown in figure 1, the coverage of the same natural person is distributed between 0.1 and 0.5, indicating that the degree ranking of the same natural person at the corresponding nodes of Twitter and Foursquare has a certain similarity. In the top 10% node-set, the same natural person coverage of Twitter and Foursquare reaches 0.5, this indicates that among the top 10% of the identity aligned nodes in the two data sets, 50% of their neighbor nodes correspond to the same natural person. In the node-set after the top 10%, the same natural person coverage is less than 0.3. The comparison results show that this degree ranking similarity is stronger in the top 10% node-set. Nodes with a higher degree are called hub nodes in the network. Such nodes usually have a higher identification degree in the network. Corresponding to the identity association scene, nodes with a higher degree have a stronger positioning effect.

**Conclusion 1.** In social networks, different nodes have different positioning functions, and the nodes with a higher degree have a strong identity association positioning function.

**Similarity analysis of adjacent nodes.** Generally, a natural person has similar friend groups in different social networks.\(^3\) Given node $v_i \in G_i$, $v_j \in G_j$ in Twitter $G_i$ and Foursquare $G_j$ social network, respectively, analyze the similarity of $i$-hop adjacent nodes belonging to the same natural person and different natural persons.\(^3\)
where $N_l(v)$ represents the set of $l$-hop adjacent nodes of node $v$. $F(V)$ represents the natural person set corresponding to node-set $V$. We draw the similarity histogram. The ordinate is the similarity interval. The abscissa is the proportion of nodes in the current similarity interval. In order to facilitate comparison, we mark the proportion of similarity of adjacent nodes of different natural persons as negative.

As shown in the abscissa positive part in the similarity comparison diagram of 1-hop adjacent nodes in Figure 2(a), the similarity of adjacent nodes belonging to the same natural person is scattered, most of which are distributed between 0 and 1, while most of the similarity of adjacent nodes not belonging to the same natural person is 0, and a few are scattered between 0.1 and 0.6. It can be seen from the comparison that the similarity of adjacent nodes belonging to the same natural person is significantly higher than that of adjacent nodes belonging to the non-same natural person.

As shown in the similarity comparison diagram of two-hop adjacent nodes in Figure 2(b), when two-hop neighbor nodes belong to the same natural person and different natural persons, most of them are distributed between 0.5 and 1.0 (as shown by the positive and negative values of the horizontal axis in Figure 2(b)), and the distinction between the same natural person and non-same natural person is not high. The location advantage of this indirect neighbor node to identity association is not apparent.

The above comparative analysis shows that in the identity association scenario, the set of one-hop adjacent nodes of the same natural person has certain similarities (i.e. the similarity between neighbor nodes belonging to the same natural person in different domains), which indirectly shows that different adjacent nodes have additional support for identity association.
Conclusion 2. For a natural person, the importance of identity association reflected by different adjacent nodes from different domains is different, and its direct neighbors support identity association more.

Proposed method

According to the above two conclusions, this article proposes a supervised learning identity association method divided into two parts: the representation learning model and the identity association model. First, according to the important connectivity role of the hub node in the relational network, we use the representation learning method to make the user node in the relational network have the location features and importance features of the node at the same time (such as the analysis of the data set in section “Data set analysis”), and then through the identity association prediction model, Use the pseudo-Siamese network to calculate the similarity between user nodes. The details are as given below.

Representation learning model. In this article, to facilitate feature representation and the calculation of final identity association similarity, we first need to represent different users in the potential space. According to the two conclusions from the analysis in the previous section, the following problems need to be solved in the process of representation learning:

1. How to characterize the positioning effect of nodes with a large degree of identity association.
2. How to characterize the different vital roles of adjacent nodes on identity association.

For problem 1, the node needs to be represented by an eigenvector, which should reflect the node’s position in the social network. For problem 2, we need to use a mechanism to calculate the importance of adjacent nodes of this node adaptively. This importance is the support of adjacent nodes for the node’s identity association and the node’s location in the network.

In order to solve the above problems, this article designs a new network representation learning method, which uses two types of vectors to represent each node: location feature and importance feature. Combining questions 1 and 2, the representation learning objectives are as follows:

Objective 1 (O1): in the same network, learn the importance feature of node $v$, which is used to calculate the support of its adjacent nodes and its location feature in the overall network structure. The higher the similarity between the location features of the current node and the location features of the aggregated adjacent nodes. This goal is used to learn the structure of social networks in two networks.

Objective 2 (O2): in the cross-domain network, the location feature similarity of different nodes belonging to the same natural person is higher than that of different nodes not belonging to the same natural person. This target is used for identity association.

In social network $G$, As shown in figure 3, we define two potential spaces: location feature potential space $\mathcal{F} \subseteq \mathbb{R}^U \times \mathbb{R}^{DF}$ and importance feature potential space $\mathcal{H} \subseteq \mathbb{R}^U \times \mathbb{R}^{DH}$, where $U$ denotes the number of nodes in network $G$, $DF$ and $DH$ denote the feature dimension of feature potential space $\mathcal{F}$ and $\mathcal{H}$, respectively. Given the user node $v_m$ in social network $G$, its location feature is expressed as $z_m^F \in \mathcal{F}$ and its importance feature is expressed as $z_m^I \in \mathcal{H}$. Given the adjacent node $v_n$ of the node $v_m$, its location feature is expressed as $z_n^F \in \mathcal{F}$, its importance feature is expressed as $z_n^I \in \mathcal{H}$.

In order to effectively integrate the location features of adjacent nodes and calculate the similarity, this article first uses the attention mechanism. It calculates the support of adjacent nodes combined with the importance feature, given the adjacent node $v_n$ of $v_m \in G$, the support $\alpha_{mn}$ of $v_n$ to $v_m$ can be expressed as follows

$$\alpha_{mn} = \frac{z_m^I \cdot z_n^T}{\sqrt{\sum_{v_i \in N_G(v_m)} (\alpha_{mn} \cdot z_n^T)}}$$

where the superscript $T$ represents the vector transpose operation. Normalize the support of all adjacent nodes.
of \( v_n \), which is used to support subsequent feature fusion. The normalized support is expressed as \( \beta_{mn} \)

\[
\beta_{mn} = \text{softmax}(\alpha_{mn}) = \text{softmax}(\frac{e^{x_m^T \cdot z_n}}{\sum_{k \in N(v_m)} e^{x_m^T \cdot z_k}})
\]

(5)

where \( N(v_m) \) represents a set of all the adjacent nodes of \( v_m \), and softmax(\( x_i \)) = \( \frac{e^{x_i}}{\sum_{j} e^{x_j}} \). Combined with the normalized support of all adjacent node location features of node \( v_m \), we calculate the aggregation feature \( e_m^F \)

\[
e_m^F = \sum_{n \in N(v_m)} \beta_{mn} \cdot e_n^F
\]

(6)

After calculating the aggregated features, we can optimize the training of node location features and importance features in representation learning. Corresponding to the target O1, the higher the similarity between the node location features and the aggregated location features of all adjacent nodes as follows

\[
p^S = - \log \left( \text{sigmoid}(z_m^T \cdot e_m^F) \right) = - \log \left( \frac{1}{1 + e^{-z_m^T \cdot e_m}} \right)
\]

(7)

Corresponding to target O2, the higher the similarity of different node location features of the same natural person in different domains as follows

\[
p^F = - \log \left( \text{sigmoid}(z_m^T \cdot z_m^F) \right) = - \log \left( \frac{1}{1 + e^{-z_m^T \cdot z_m}} \right)
\]

(8)

where activation function is sigmoid(\( x \)) = \( 1/(1 + e^{-x}) \). Equations (7) and (8) above convert the two objectives of representation learning into binary classification problems, respectively.

In order to avoid trivial solutions, we introduce the negative sampling term in the training and optimization process of the above formula. In the optimization process for target O1, we introduce the case of non-adjacent nodes, and in the process for target O2, we introduce the case of non-identical natural persons. Therefore, in the process of calculating node aggregation feature \( e_m^F \), if all the nodes participating in the calculation meet the condition of aggregating the features of all adjacent nodes, that is, \( n \in N(v_m) \), then the expectation is \( p^S \rightarrow 1 \); otherwise, if \( n \notin N(v_m) \), then the expectation is \( p^S \rightarrow 0 \). If \( v_m, v_n \) belong to the same natural person, then the expectation is \( p^F \rightarrow 1 \); if \( v_m, v_n \) do not belong to the same natural person, then \( p^F \rightarrow 0 \).

For the binary classification problem of calculation formulas (7) and (8), the binary cross-entropy is used as the loss function for optimization

\[
\mathcal{L}_1^S = - \frac{1}{K_S} \sum_{i=1}^{K_S} y_i^S \cdot \log(p_i^S) + (1 - y_i^S) \cdot \log(1 - p_i^S)
\]

(9)

\[
\mathcal{L}_2^F = - \frac{1}{K_F} \sum_{i=1}^{K_F} y_i^F \cdot \log(p_i^F) + (1 - y_i^F) \cdot \log(1 - p_i^F)
\]

(10)

\( \mathcal{L}_1^F \) is the loss function value representing the learning target O1, and \( \mathcal{L}_2^F \) is the loss function value representing the learning target O2. The two social networks \( G_i, G_j \) contain two O1 loss functions \( \mathcal{L}_1^{F_1}, \mathcal{L}_1^{F_2} \) and one O2 loss function in the actual training process. Finally, the total loss of learning is

\[
\mathcal{L}^E = \mathcal{L}_1^{F_1} + \mathcal{L}_1^{F_2} + \mathcal{L}_2^F
\]

(11)

where \( \mathcal{L}^E \) is the total loss of learning in the potential space of different users across two network domains, and this total loss function can be extended to multiple network domains.

**Identity association model.** Given the social network \( G_i, G_j \), as shown in figure 4, there are users \( v_p \in G_i \) and \( v_q \in G_j \). After feature learning through the representation learning model, we get the location feature \( z_p^F, z_q^F \) and the importance feature \( z_p^H, z_q^H \). The location features are aggregated by the location features of all neighbor nodes of the current node. However, the support of all neighboring nodes to the current node is reflected by the importance characteristics, and the support of the same node to other adjacent nodes is also different. Therefore, the location feature reflects the environmental information of adjacent nodes, while the importance feature reflects the unique support relationship. Compared with location features, importance features have more robust discrimination and are more suitable for identity association scenarios. In this article, the importance feature is input into the identity association model to predict the final identity association.

In this article, because the number of training set samples is small, and the identity association belongs to the nearest similarity matching problem, we use a pseudo-Siamese Network for transformation to avoid the result of overfitting. Moreover, the pseudo-Siamese Network itself is used to measure the similarity between the two inputs. In order to learn the linear relationship between features, we use a linear function in the activation function.
The input importance feature dimensions may differ for users from two different social networks. First, the two importance features are transformed into the same dimension through a perceptron layer. We define weight matrix $W^i \in \mathbb{R}^{D_f \times t_i}$, $W^j \in \mathbb{R}^{D_f \times t_j}$ and bias vectors $b^i \in \mathbb{R}^{t_i}$, $b^j \in \mathbb{R}^{t_j}$, respectively, where $t$ represents the dimension of the output vector. $z^H_q$, $z^H_p$ are used as the input vectors, and the output intermediate process vectors are $h_p$, $h_q \in \mathbb{R}^t$ after passing through the single-layer perceptron. The process is as follows:

$$h_p = \tanh \left( W^i : z^H_p + b^i \right)$$

$$h_q = \tanh \left( W^j : z^H_q + b^j \right)$$

Finally, the identity association similarity of $v_p$ and $v_q$ is calculated as follows:

$$p = \tanh \left( (h_p \times h_q)^T \cdot (h_p \times h_q) \right)$$

where $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$. Since the result of $(h_p \times h_q)^T \cdot (h_p \times h_q)$ is great than 0, similarly, we use the activation function $\tanh$ for $[0, 1]$ normalization.

In order to avoid the trivial solution, we introduce the negative sampling term in the training optimization process of the above formula. If $v_p, v_q$ are the same natural person, then $p \to 1$; if $v_p, v_q$ are not the same natural person, then $p \to 0$. The binary cross-entropy function is used as the loss function in negative sampling:

$$L^C = - \frac{1}{K} \sum_{i=1}^{K} \left( y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p_i) \right)$$

where $x$ represents the sample and $y$ represents the output value from two to label 0 or 1, $p(y)$ is the probability of the actual output label $y$, and $k$ represents the total number of samples. If the predicted value $p(y)$ approaches to 1, the loss function value should approach 0. On the contrary, if the predicted value $p(y)$ tends to 0, the loss function value should be huge.

**Method overview**

Algorithm 1 gives the pseudo-code of the algorithm in this article.

Note that user identity association usually involves more than two network domains in practical applications. Most of the previous research and engineering development work focused on the user identity association of the two domains. In the application scenario of user identity association across more than two network domains, we usually need to retrain the model due to different user attributes, generated content, or network relationship dimensions in different network domains. Our model only considers the importance of the degree of nodes and neighbor nodes and does not involve user attributes and generated content. Therefore, our method can be easily extended to application scenarios such as user entity identity association across multiple network domains, IoT relationship discovery, behavior traceability, and resources surveying mapping.

**Experiments**

First, we evaluate the user entity association method in sections “P@N” and “H@N” in a cross-domain environment, then evaluate the effect of different training ratios in section “Comparison of effects of different training ratios,” and finally, evaluate the comparison between the effect in section “Comparison with other models” and other models on the same data set.
Evaluation method

We use $P@N$ and $H@N$ to evaluate the identity association model in this article. Among them, $P@N$ is used to evaluate the ranking proportion of nodes that meet similar requirements in all nodes. $H@N$ is used to evaluate the ranking proportion of association similarity after node association, which is the sum of the number of users successfully associated. In essence, the final output of identity association of entity users in a cross-domain network is the probability that users belong to the same natural person. In the cross-domain network environment, the higher the ranking of this probability, the greater the probability of belonging to the same natural person. The traditional method uses accuracy and recall to evaluate the effect of identity association. In essence, it is to find the similarity threshold through the model. When the output similarity is greater than this threshold, it belongs to the same natural person. The disadvantage of this method is that there may be many similarity outputs greater than the threshold; however, the larger the output value (i.e. the higher the similarity ranking), the greater the probability that two or more of them belong to the same natural person. Traditional methods ignore the key and strong identification role of ranking similarity values output by the model in entity identity association.

$P@N$. In many user entity identity association methods, a mature and widely used evaluation index is to compare the first $n$ candidates of identity association. Given two social networks $G_i, G_j$, $|CorrUser@N|$ represents the number of similarities in top-$N$ when $G_i$ is associated with $G_j$, $|CorrUser@N|$ represents the number of similarities in top-$N$ when $G_j$ is associated with $G_i$, $|UnMappedUsers|$ indicates the number of users to be associated. In cross-domain networks, a similar node ranking calculation method $P@N$ is as follows

$$P@N = \frac{|CorrUser@N| + |CorrUser@N|}{|UnMappedUsers| \times 2}$$

As the association results $P@N$ indicate the proportion of identities ranked in the top $n$ identity association similarity. The larger the value, the better the identity association effect.

$H@N$. Evaluation method $P@N$ ignores the sorting relationship in the first $n$ candidates. In this article, $H@N$ by comparing the top-$N$ identity association results, the identity association effect is evaluated. The higher the identity association similarity ranking, the higher the value of $H@N$. The function $h(x)$ represents the sorting effect of association results in top-$N$, as follows

$$h(x) = \frac{N - (hit(x) - 1)}{N}$$

Here, $H@N$ is the proportion of the sum of nodes satisfying identity association in all $N$ nodes. The higher the proportion, the better the effect of identity association. For example, given a test set $X = \{x_1, x_2, \ldots, x_k\} (k \in N)$, $j, j \in k$, when $x_i$ is the last node that meets the identity association conditions, then $H@N = \frac{\sum_{i=1}^{k} h(x_i) / k}$ represents the sum of the number of nodes that meet the requirements of the identity association model, that is, the ranking proportion of nodes that meet the requirements of the identity association model.

$P@N$ and $H@N$ are used to evaluate the effect of identity association. The similarity ranking $N$ is the identity association threshold in the actual identity association prediction. When the identity association similarity ranking appears in the top $N$, it is generally considered that it may belong to the same identity.

Comparison of effects of different training ratios

In order to evaluate the model more comprehensively in this article, we set different proportions of training samples for cross-domain user identity association similarity and cross-domain user identity association effect and use Twitter–Foursquare data set to analyze the training effects of methods $P@N$ and $H@N$ in this article at different training ratios.

First, we set the training proportion of the number of users to be associated to 0.1–0.9. Under different candidate sets, we analyze the ranking of similar nodes and association similarity of cross-domain entity users. When analyzing the ranking $P@N$ of user similar nodes, different similar node rankings are set to the top 10, 30, and 50, respectively. Because the ranking analysis requirements of $H@N$ are more stringent than those of $P@N$ and require smaller analysis granularity, so the ranking setting of 100 is added; that is, the association similarity ranking of different nodes is set to the top 10, 30, 50, and 100, respectively. The test results are shown in Tables 2 and 3 and Figure 5. The details of algorithm execution can be easily observed from the table, and the performance trend of the algorithm when
the marked number changes can be easily judged from the figure.

The results in Table 2 show the comparison of P@N evaluation performance when the ranking of similar nodes is 10, 30, and 50, respectively. It can be noted that when the node similarity ranking value is the same, the proportion of identity association ranking in the test set increases with the increase of the proportion of training data set from 0.1 to 0.9. When the proportion of candidate sets is the same, as the node similarity ranking value gradually increases from 10 to 30, 50, the proportion of identity association ranking in the test set increases. According to the results of many tests, when the training proportion is 90%, and the similar nodes rank in the top 50, the number of nodes meeting the requirements of the identity association test accounts for 55.9%, and the effect is the best, as shown in Figure 5(a).

In Table 3 and Figure 5(b), we give a similarity evaluation method for user entity identity association in a cross-domain environment and compare the performance of H@N when similar nodes rank top 10, 30, 50, and 100, respectively. In the range of 0.1–0.9 of the proportion of the whole candidate set, the evaluation
effect trend of each association similarity ranking shows obvious consistency; that is, the proportion of each association similarity ranking shows an increasing trend, which clearly shows that the representation learning method proposed in this article improves the effectiveness of entity identity association. When the metric is taken as 10, the increase of evaluation value is not obvious. When the metric is taken as the first 100, the increase of evaluation value is the most obvious and finally reaches 0.489.

Comparison with other models

We tested EUIA\(^{22}\) and APAN\(^{25}\) algorithms on the same test set, all based on network-embedding learning and using only network structure. EUIA is an identity association method based on LINE (large-scale information network embedding),\(^{29}\) and APAN is an identity association method based on DeepWalk\(^{20}\) network-embedding. They only use the network structure to analyze the association of network entity identity. Through comparative analysis, it is found that our method performs well on the whole.

First, we compare IFN-UIL, EUIA, and APAN algorithms when the proportion of similar nodes is P@50, the user entity association similarity is H@50, and we observe that when \(N = 50\), the accuracy of the ranking of similar nodes of IFN-UIL is 45.5% and 30% higher than that of EUIA and APAN, respectively, and the ranking of user entity association similarity is about 30% and 25% higher than that of EUIA and APAN, respectively. The main reasons are as follows:

1. EUIA uses the first-order LINE network-embedding identity association method. This method only models the local adjacency of each node but ignores the global connection characteristics of nodes in the network. APAN uses truncated RandomWalk sequences to learn to embed and capture local and global structural attributes. However, the above two methods ignore the positioning role of the Hub Node in the whole network.

2. We propose a new network representation learning method in which each node is represented by location feature and importance feature. We finally selected the node importance feature with a better effect to calculate the identity association similarity of cross-domain network entity users. The output result is the ranking of the proportion of identity association similarity. This representation method of hit accuracy has finer granularity than the threshold discrimination method used by EUIA and APAN, so the accuracy of our results is better.

This conclusion can be verified by Figure 6. Figures 5(b) and 6(a) describe the performance changes of methods IFN-UIL, EUIA, and APAN, respectively. With the increase in the number of iterations, it can be found that there is a large gap in the results of various methods. IFN-UIL has achieved better results and obvious advantages than EUIA and APAN for the above reasons.

Conclusion

This article proposes an entity identity association model IFN-UIL based on supervised learning. Based on the association positioning function of hub nodes, this model is used to solve the probabilistic alignment problem of network entity identity in cross-network domain scenarios. IFN-UIL represents each node in multiple cross-domain networks through two low-dimensional vectors: node location feature and importance feature. We select the node importance feature...
which has a better effect on identity association for the final identity association prediction. This article uses real data sets to verify and evaluate the performance of the proposed IFN-UIL model. It verifies the effectiveness of our proposed model in the application scenario of entity identity association in cross-domain networks. The experimental results show that the effect of IFN-UIL has certain advantages in the situation of using only network structure information.

The association analysis method of inter-entity relationship based on the association positioning function of hub nodes proposed in this article can be easily extended to the fields of “human-thing” relationship discovery of IoT, network abnormal behavior analysis, cyberspace resources surveying mapping, and so on.

IFN-UIL has two limitations. First, since the IFN-UIL model depends on the results of representation learning, when the network structure changes, such as the addition and deletion of nodes, the model needs to be retrained. Second, because the node embedding vector generated by IFN-UIL is related to the network structure, the accuracy will be low when the network structure of the actual cross-domain application scenario is entirely different. Our subsequent work will try to solve the above problems and improve the efficiency and scalability of the model.

Declaration of conflicting interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Key R&D Program of China under Grant 2020YFB1708600.

ORCID iDs
Mingcheng Gao https://orcid.org/0000-0002-5246-1162
Ruiheng Wang https://orcid.org/0000-0001-6833-9551
Hongliang Zhu https://orcid.org/0000-0003-2448-2027

References
1. Farseev A, Nie L, Akbari M, et al. Harvesting multiple sources for user profile learning: a big data study. In: Proceedings of the 5th ACM international conference on multimedia retrieval (ICMR), Shanghai, China, 23–26 June 2015, pp.235–242. New York: ACM.
2. Song X, Nie L, Zhang L, et al. Multiple social network learning and its application in volunteerism tendency prediction. In: Proceedings of the 38th ACM SIGIR, Santiago, Chile, 9–13 August 2015, pp.213–222. New York: ACM.
3. Wu X, Hu Z, Fu X, et al. Social network de-anonymization with overlapping communities: analysis, algorithm and experiments. In: Proceedings of the 37th IEEE international conference on computer communications (INFOCOM), Honolulu, HI, 16–19 April 2018. New York: IEEE.
4. Yao L, Sheng QZ, Falkner NJG, et al. ThingsNavi: finding most-related things via multi-dimensional modeling of human-thing interactions, ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2014, https://www.researchgate.net/publication/286753251_ThingsNavi_Finding_Most-Related_Things_via_Multi-Dimensional_Modeling_of_Human-Thing_Interactions
5. Zhou Y, Xu Q, Luo XY, et al. Research on definition and technological system of cyberspace surveying and mapping. Comput Sci 2018; 45: 1–4.
6. Wen YT, Lei PR, Peng WC, et al. Exploring social influence on location-based social networks. In: Proceedings of the IEEE international conference on data mining (ICDM), Shenzhen, China, 14–17 December 2014, pp.1043–1048. New York: IEEE.
7. Li J, Cai Z, Yan M, et al. Using crowdsourced data in location-based social networks to explore influence maximization. In: Proceedings of the 35th IEEE international conference on computer communications (INFOCOM), San Francisco, CA, 10–14 April 2016. New York: IEEE.
8. Zhou J, Fan J, Wang J, et al. Cost-efficient viral marketing in online social networks. World Wide Web 2019; 22: 2355–2378.
9. Zafarani R and 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 (KDD 2013), Chicago, IL, 11–14 August 2013, pp.41–49. New York: ACM.
10. Shen Y and Jin H. Controllable information sharing for user accounts linkage across multiple online social networks. In: Proceedings of the 23rd ACM international conference on information and knowledge management (CIKM 2014), Shanghai, China, 3–7 November 2014, pp.381–390. New York: ACM.
11. Liu S, Wang S, Zhu F, et al. HYDRA: large-scale social identity linkage via heterogeneous behavior modeling. In: Proceedings of the 2014 ACM SIGMOD international conference on management of data (SIGMOD 2014), Snowbird, UT, 22–27 June 2014, pp.51–62. New York: ACM.
12. Zhang H, Kan MY, Liu Y, et al. Online social network profile linkage. In: Proceedings of the 10th AIRS conference on information retrieval technology, 2014, pp.197–208, https://www.comp.nus.edu.sg/~kanny/papers/airs2014.pdf
13. Zheng R, Li J, Chen H, et al. A framework for authorship identification of online messages: writing-style features and classification techniques. J Assoc Inform Sci Technol 2006; 57(3): 378–393.
14. Almishari M and Tsudik G. Exploring linkability of user reviews. In: Proceedings of the European symposium on research in computer security, 2012, https://petsymposium.org/2012/papers/hotpets12-6-yelp.pdf
15. Huang C, Ni J, Lu R, et al. Exploring anonymous user reviews: linkability analysis based on machine learning. In: Proceedings of the 2019 IEEE global communications conference (GLOBECOM), Waikoloa, HI, 9–13 December 2019. New York: IEEE.

16. Zhou J and Fan J. TransLink: user identity linkage across heterogeneous social networks via translating embeddings. In: Proceedings of the IEEE INFOCOM 2019—IEEE conference on computer communications, Paris, 29 April–2 May 2019. New York: IEEE.

17. Guo Y, Liu Z, Zeng Y, et al. Preserving privacy for hubs and links in social networks. In: Proceedings of the 2018 international conference on networking and network applications, Xi’an, China, 12–15 October 2018. New York: IEEE.

18. Tan S, Guan Z, Cai D, et al. Mapping users across networks by manifold alignment on hypergraph. In: Proceedings of the 28th conference on artificial intelligence (AAAI), 2014, https://ojs.aaai.org/index.php/AAAI/article/view/8720

19. Zhou F, Liu L, Zhang K, et al. DeepLink: a deep learning approach for user identity linkage. In: Proceedings of the 37th IEEE international conference on computer communications (INFOCOM), Honolulu, HI, 16–19 April 2018. New York: IEEE.

20. Perozzi B, Al-Rfou R and Skiena S. DeepWalk: online learning of social representations. In: Proceedings of the 20th ACM international conference on knowledge discovery and data mining (SIGKDD), 2014, pp.701–710, https://arxiv.org/pdf/1403.6652.pdf

21. Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data. In: Proceedings of the 27th annual conference on neural information processing systems (NIPS), 2013, pp.2787–2795, https://proceedings.neurips.cc/paper/2013/file/1ece7a77928ca8133fa24680a88d29-Paper.pdf

22. Miao Q, Wang L, Duan D, et al. Embedding based cross-network user identity association technology. In: Proceedings of the 2019 3rd international conference on digital signal processing, Jeju, South Korea, 24–26 February 2019. New York: ACM.

23. Sun B, Zhang W and Si CX. Social network user identity association and its analysis. J Beijing Univ Post Telecommun 2020; 43(1): 126–132.

24. Li Z, Zheng W, Lin X, et al. TransN: heterogeneous network representation learning by translating node embeddings. In: Proceedings of the 2020 IEEE 36th international conference on data engineering (ICDE), Dallas, TX, 20–24 April 2020. New York: IEEE.

25. Wang S, Li X, Ye Y, et al. Anchor link prediction across attributed networks via network embedding. Entropy 2019; 21(3): 254.

26. Yao L and Sheng QZ. Exploiting latent relevance for relational learning of ubiquitous things. In: Proceedings of the IEEE international conference on data mining (ICDM 2013), Dallas, TX, 7–10 December 2013. New York: IEEE.

27. Mu X, Zhu F, Lim EP, et al. User identity linkage by latent user space modelling. In: Proceedings of the 22nd ACM SIGKDD international conference, 2016, https://www.kdd.org/kdd2016/papers/files/rrp0741-muAemb.pdf

28. Tang J, Qu M, Wang M, et al. LINE: large-scale information network embedding. In: Proceedings of the international conference on world wide web, 2015, https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/frp0228-Tang.pdf

29. Zhang J and Yu PS. Integrated anchor and social link predictions across social networks. In: Proceedings of the 24th international joint conference on artificial intelligence, Buenos Aires, Argentina, July 2015, pp.2125–2132, http://shichuan.hit.edu.cn/topic/Information%20Fusion/2015.%20Integrated%20anchor%20and%20social%20link%20predictions%20across%20social%20networks.pdf

30. Li Y, Su Z, Yang J, et al. Exploiting similarities of user friendship networks across social networks for user identification. Inform Sci 2020; 506: 78–98.

31. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Adv Neur Inform Process Syst 2017; 3762: 5998–6008.

32. Liu L, Li X, Cheung WK, et al. Structural representation learning for user alignment across social networks. IEEE Trans Knowl Data Eng 2020; 32(9): 1824–1837.