

Singapore Management University Institutional Knowledge at Singapore Management University

Research Collection School Of Information Systems

School of Information Systems

11-2018

Unsupervised user identity linkage via factoid embedding

Wei XIE

Singapore Management University, weixie@smu.edu.sg

Xin MU

Nanjing University

Roy Ka Wei LEE

Singapore Management University, roylee@smu.edu.sg

Feida ZHU


Singapore Management University, fdzhu@smu.edu.sg

Ee-peng LIM

Singapore Management University, eplim@smu.edu.sg

DOI: <https://doi.org/10.1109/ICDM.2018.00182>

Follow this and additional works at: https://ink.library.smu.edu.sg/sis_research

 Part of the [Databases and Information Systems Commons](#), and the [Theory and Algorithms Commons](#)

Citation

XIE, Wei; MU, Xin; LEE, Roy Ka Wei; ZHU, Feida; and LIM, Ee-peng. Unsupervised user identity linkage via factoid embedding. (2018). *2018 IEEE International Conference on Data Mining ICDM 2018: Singapore, November 17-20: Proceedings*. 1338-1343. Research Collection School Of Information Systems.

Available at: https://ink.library.smu.edu.sg/sis_research/4258

This Conference Proceeding Article is brought to you for free and open access by the School of Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email libIR@smu.edu.sg.

Unsupervised User Identity Linkage via Factoid Embedding

Wei Xie*, Xin Mu[†], Roy Ka-Wei Lee*, Feida Zhu* and Ee-Peng Lim*

*Living Analytics Research Centre

Singapore Management University, Singapore

Email: {weixie, roylee.2013, fdzhu, eplim}@smu.edu.sg

[†]National Key Laboratory for Novel Software Technology

Nanjing University, China

Email: mux@lamda.nju.edu.cn

Abstract—User identity linkage (UIL), the problem of matching user account across multiple online social networks (OSNs), is widely studied and important to many real-world applications. Most existing UIL solutions adopt a supervised or semi-supervised approach which generally suffer from *scarcity of labeled data*. In this paper, we propose **Factoid Embedding**, a novel framework that adopts an unsupervised approach. It is designed to cope with different profile attributes, content types and network links of different OSNs. The key idea is that each piece of information about a user identity describes the real identity owner, and thus distinguishes the owner from other users. We represent such a piece of information by a *factoid* and model it as a triplet consisting of *user identity*, *predicate*, and an *object* or another *user identity*. By embedding these factoids, we learn the user identity latent representations and link two user identities from different OSNs if they are close to each other in the user embedding space. Our **Factoid Embedding** algorithm is designed such that as we learn the embedding space, each embedded factoid is “translated” into a motion in the user embedding space to bring similar user identities closer, and different user identities further apart. Extensive experiments are conducted to evaluate **Factoid Embedding** on two real-world OSNs data sets. The experiment results show that **Factoid Embedding** outperforms the state-of-the-art methods even without training data.

Index Terms—user identity linkage, factoid embedding, network embedding

I. INTRODUCTION

Motivation. Increasingly, people are using multiple online social networks (OSNs) to meet their communication and relationship needs¹. The rise of users using multiple OSNs motivates researchers to study User Identity Linkage (UIL), the problem of linking user accounts from different OSNs belonging to the same person. Tackling UIL is imperative to many applications, particularly user profiling and recommender systems.

User Identity Linkage Problem. The UIL problem has been widely studied and is usually formulated as a classification problem, i.e. to predict whether a pair of user identities from different OSNs belong to the same real person [1]. There are many supervised and semi-supervised methods proposed to address UIL but they could not perform well when there is

a *scarcity of labeled data*. One possible way to obtain labeled matching user accounts is to recruit users to manually identify them. Such an approach is very costly and time consuming. In this research, we therefore aim to solve the UIL problem using an unsupervised approach.

User Identity Linkage Problem in Unsupervised Setting.

We formulate the UIL problem in unsupervised setting as follows. Let u be a user identity in an OSN which belongs to a real person p . Let $\mathbf{o}_u = [\mathbf{o}_{u,1}, \dots, \mathbf{o}_{u,d}]$ denote a set of data objects associated with u . These objects include username, screen name, profile image, profile description, posts, etc.. We denote an OSN as $\mathcal{G} = (\mathcal{U}, \mathcal{O}, \mathcal{E})$, where $\mathcal{U} = \{u_1, \dots, u_N\}$ is the set of user identities, $\mathcal{O} = \{\mathbf{o}_{u_1}, \dots, \mathbf{o}_{u_N}\}$ is the set of corresponding data objects, and $\mathcal{E} \subseteq \mathcal{U} \times \mathcal{U}$ is the set of links in the network. Thus, given two OSNs $\mathcal{G}^s = (\mathcal{U}^s, \mathcal{O}^s, \mathcal{E}^s)$ (source) and $\mathcal{G}^t = (\mathcal{U}^t, \mathcal{O}^t, \mathcal{E}^t)$ (target), without any known matched user pairs between \mathcal{G}^s and \mathcal{G}^t , the objective is to return a user u^t in target OSN, for every user u^s in the source OSN, such that the user pair (u^s, u^t) most likely belongs to the same real person p .

Research Objectives and Contributions. In this paper, we propose Factoid Embedding, a novel framework that links user identities across multiple OSNs through the use of a network embedding approach. The key idea behind Factoid Embedding is that despite the heterogeneity in information from multiple OSNs, each piece of information about a user identity describes the person who owns it, and thus help to distinguish the person from others. The more information we have, the closer we get to know about the real person. Specifically, we model each piece of information as a factoid, which is a triplet consisting of *user identity*, *predicate* and an *object* or another *user identity* (as shown in Table I). Embedding these factoids provides the unifying structure to represent the heterogeneous information and data types. Figure 1 shows the framework of Factoid Embedding. Firstly, we generate the factoids from information gathered from different OSNs. Next, we embed heterogeneous objects (e.g. names, texts, and images) into their respective embedding spaces (e.g., names will be embedded into the name embedding space, etc.) incorporating the similarity measures that contribute to matching user identities. Note that when embedding the

¹www.si.umich.edu/news/more-adults-using-multiple-social-platforms-survey-finds

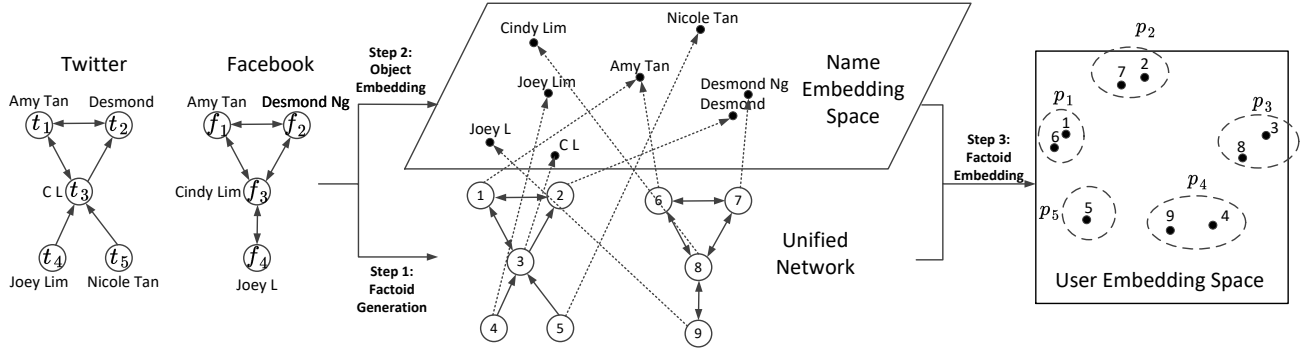


Fig. 1: Framework of Factoid Embedding

heterogeneous objects, we leverage on external and prior knowledge such that if two objects are similar, their embedding vectors will be close to each other in the object embedding space. For example, the user names *Desmond* and *Desmond Ng* are similar and therefore, the embedding vectors of the two names will be close to each other in the name embedding space. Finally, leveraging the factoids’ triplet structure, we project the various object embeddings into the user embedding space. Essentially, the vectors in the user embedding space represent the user identities, and through iterations of object embedding projections, the user identities that share many similar factoids will be “pushed” closer to one another in the user embedding space.

Overall, this paper improves the state-of-the-art by making the following contributions:

- We propose a novel unsupervised method called **Factoid Embedding** to link user identities from different OSNs. Our proposed method is able to integrate heterogeneous information using the triplet structure of factoids and object embeddings. To the best of our knowledge, this is the first work which embeds heterogeneous information to address the UIL problem in an unsupervised manner.
- We conduct extensive experiments on three real-world OSNs, namely, Twitter, Facebook and Foursquare, to evaluate our proposed model. The experiment results show that Factoid Embedding outperforms the state-of-the-art methods. It can even outperform some of the existing supervised methods which are given labeled matching user pairs for training.

II. PROPOSED SOLUTION

We propose an integrated three-step solution which is illustrated in Figure 1. In the first step, we use the information from \mathcal{G}^s and \mathcal{G}^t to generate a set of factoids. Next, we embed the heterogeneous data objects (e.g. names, text and images) into their respective embedding spaces. Finally, we learn and project the user identities and their links into user embedding space using factoid embedding, a process where we integrate the object embeddings and factoids.

TABLE I: Examples of Generated Factoids

Twitter	Facebook
$\langle 1, has_name, Amy\ Tan \rangle$	$\langle 6, has_name, Amy\ Tan \rangle$
$\langle 2, has_name, Desmond \rangle$	$\langle 7, has_name, Desmond\ Ng \rangle$
$\langle 3, has_name, C\ L \rangle$	$\langle 8, has_name, Cindy\ Lim \rangle$
$\langle 4, has_name, Joey\ Lim \rangle$	$\langle 9, has_name, Joey\ L \rangle$
$\langle 5, has_name, Nicole\ Tan \rangle$	$\langle 6, follows, 7 \rangle$
$\langle 1, follows, 2 \rangle$	$\langle 7, follows, 6 \rangle$
$\langle 2, follows, 1 \rangle$	$\langle 6, follows, 8 \rangle$
$\langle 1, follows, 3 \rangle$	$\langle 8, follows, 6 \rangle$
$\langle 3, follows, 1 \rangle$	$\langle 7, follows, 8 \rangle$
$\langle 3, follows, 2 \rangle$	$\langle 8, follows, 7 \rangle$
$\langle 4, follows, 3 \rangle$	$\langle 8, follows, 9 \rangle$
$\langle 5, follows, 3 \rangle$	$\langle 9, follows, 8 \rangle$

A. Factoid Generation

To integrate the heterogeneous user attribute/content objects and their user-user link information in \mathcal{G}^s and \mathcal{G}^t , we first combine and represent the information in an unified network. In this unified network, every user identity u_i (from source or target network) is represented as a new user node with a unique ID and every data object is represented as a data node (as illustrated in the step 1 in Figure 1). We then represent a user-object association and a user-user link as an *user-object factoid* and an *user-user factoid* respectively. A user-object factoid $\langle u_i, pred, o \rangle$ has *pred* denoting the associated attribute predicate, and *o* denoting a data object. Each user-object factoid provides us a description about u_i . For example in Figure 1, factoid $\langle 1, has_name, Amy\ Tan \rangle$ conveys the information that the u_1 has name “Amy Tan”. Next, we use another set of predicates to represent user-user links. For example, for Twitter, an user identity may “follows” another user identity. As such, we represent *follows* as a predicate and let $\langle u_i, follows, u_j \rangle$ denote a user-user factoid with the predicate “follows”. For instance, factoid $\langle 1, follows, 3 \rangle$ tells us u_1 follows u_3 . Table I presents all the factoids generated from the two OSNs in Figure 1. In the following, we shall elaborate the embeddings of objects followed by that of user-object and user-user factoids.

B. Object Embedding

Although the factoids generated in the previous step is able to represent the different information types in a unified network, it still has to address the issue of comparing data objects of attributes used for linking user identities. For example, the factoids in row 2 of Table I do not explicitly tell us that “Desmond” and “Desmond Ng” are similar names. Instead, it only tell us that they are non-identical. Therefore, in this step we embed these heterogeneous objects taking advantage of similarity knowledge about the objects. For example, suppose two user identities sharing similar attribute objects are more likely to belong to the same person. We will then embed the objects such that similar objects are closer in the object embedding space (see step two in Figure 1, where similar names are closer in the name embedding space).

We first let O_{pred} denote all the data objects for certain predicate $pred$, i.e. $O_{pred} = \{o \in \mathcal{F}_{pred}\}$. For instance in Figure 1, $O_{has_name} = \{\text{“Amy Tan”, “Desmond”, “C L”, “Joey Lim”, “Nicole Tan”, “Desmond Ng”, “Cindy Lim”, “Joey L”}\}$. For each user-object predicate $pred$, we construct a similarity matrix S^{pred} in which each element $S_{i,j}^{pred} \in [-1, 1]$ measures the similarity between two objects $o_i, o_j \in O_{pred}$. $S_{i,j}^{pred} = 1$ when o_i and o_j are identical, and $= -1$ when o_i and o_j are completely different. There are a few ways to measure similarities between objects. For example, Jaro-Winkler distance [2] has been used to measure the similarity between two names, and deep learning techniques can help us measure how similar two profile images are. In the experiment section, we will elaborate the similarities between different types of data objects in more details.

For each data object o , we define \mathbf{v}_o to be the embedding vector of o . To learn object embeddings, we define the objective function in Equation 1. This function aims to keep the embedding vectors of similar data objects to be close to each other in the object embedding space.

$$error_{pred} = \sum_{i,j} (\mathbf{v}_{o_i}^\top \mathbf{v}_{o_j} - S_{i,j}^{pred})^2 \quad (1)$$

where $\{\mathbf{v}_o\}_{o \in O_{pred}}$ are the object embedding vectors, and S^{pred} is the given similarity matrix. We learn $\{\mathbf{v}_o\}_{o \in O_{pred}}$ by minimizing $error_{pred}$.

Ideally, $\{\mathbf{v}_o\}_{o \in O_{pred}}$ would preserve all the information in the similarity matrix S^{pred} leading to $error_{pred} = 0$. More importantly, because $S_{i,i}^{pred} = 1$, $\mathbf{v}_{o_i}^\top \mathbf{v}_{o_i} = \|\mathbf{v}_{o_i}\|_2^2$ will be close to 1, i.e. all the embedding vectors $\{\mathbf{v}_o\}$ are near to the surface of a unit hypersphere. It means $\cos(\mathbf{v}_{o_i}, \mathbf{v}_{o_j}) \approx \mathbf{v}_{o_i}^\top \mathbf{v}_{o_j} \approx S_{i,j}^{pred}$. Therefore, if o_i and o_j are similar, \mathbf{v}_{o_i} and \mathbf{v}_{o_j} will be close to each other in the embedding space. Figure 1 illustrates how the names are embedded in the name embedding space. We can see that “Desmond Ng” is close to “Desmond”, but far from “C L” in the embedding space.

In practice, the similarity matrix S^{pred} may be huge, i.e. $O((|\mathcal{U}^s| + |\mathcal{U}^t|)^2)$. In order to speed up learning, we can just focus on the similar object pairs. By employing block-techniques such as inverted index and Locality-Sensitive

Hashing (LSH) we can build a sparse similarity matrix S^{pred} . Afterwards, stochastic gradient descent is applied to minimize $error_{pred}$.

C. Factoid Embedding

In this step, we learn user identities’ latent representations by embedding the generated user-object and user-user factoids.

We let \mathcal{U}^a denote the set of all user identities in the unified network (i.e. $\mathcal{U}^a = \mathcal{U}^s \cup \mathcal{U}^t$) and \mathcal{F}_{pred} denote the set of factoids with a predicate $pred$, e.g. $\mathcal{F}_{follows} = \{\langle u_i, follows, u_j \rangle\}$, $\mathcal{F}_{has_name} = \{\langle u_i, has_name, o \rangle\}$. Suppose we have d types of user-object predicates, i.e. $\{pred_1, \dots, pred_d\}$.

For each user-object factoid in \mathcal{F}_{pred} , we define its probability as follows.

$$prob(u_i, pred, o) = \frac{\exp(\mathbf{v}_{u_i}^\top \cdot \phi_{pred}(\mathbf{v}_o))}{\sum_{u' \in \mathcal{U}^a} \exp(\mathbf{v}_{u'}^\top \cdot \phi_{pred}(\mathbf{v}_o))} \quad (2)$$

where \mathbf{v}_{u_i} is the embedding vector of user identity u_i , \mathbf{v}_o is the embedding vector of data object o , and ϕ_{pred} is a projection function which maps \mathbf{v}_o to the user embedding space. Note that we have learned \mathbf{v}_o in the object embedding step. Particularly, we impose such a constraint on ϕ_{pred} that it is a Lipschitz continuous function, i.e. there is a constant C such that $|\phi_{pred}(\mathbf{x}) - \phi_{pred}(\mathbf{y})| < C \cdot \|\mathbf{x} - \mathbf{y}\|$ for any \mathbf{x} and \mathbf{y} in the space. In other words, if two objects are similar i.e. $\mathbf{v}_{o_i} \approx \mathbf{v}_{o_j}$ then their projections will be close to each other i.e. $\phi_{pred}(\mathbf{v}_{o_i}) \approx \phi_{pred}(\mathbf{v}_{o_j})$. In this work, we set $\phi_{pred}(\mathbf{v}_o)$ as a linear function, i.e. $\phi_{pred}(\mathbf{v}_o) = \mathbf{W}_{pred} \cdot \mathbf{v}_o + \mathbf{b}_{pred}$, where \mathbf{W}_{pred} and \mathbf{b}_{pred} are unknown parameters, and \mathbf{W}_{pred} ’s norm $\|\mathbf{W}_{pred}\|$ is limited. We leave other non-linear choices of $\phi_{pred}(\mathbf{v}_o)$ for future work. Given all the user-object factoids in \mathcal{F}_{pred} , i.e. $\{\langle u_i, pred, o \rangle\}$, we define the following objective function.

$$f(\mathcal{F}_{pred}) = \sum_{\langle u_i, pred, o \rangle \in \mathcal{F}_{pred}} \log(prob(u_i, pred, o)) \quad (3)$$

Similarly, for each user-user factoid in $\mathcal{F}_{follows}$, we define its probability as follows.

$$prob(u_i, follows, u_j) = \frac{\exp(\mathbf{v}_{u_i}^\top \cdot \phi_{follows}(\mathbf{v}_{u_j}))}{\sum_{u' \in \mathcal{U}^a} \exp(\mathbf{v}_{u'}^\top \cdot \phi_{follows}(\mathbf{v}_{u_j}))} \quad (4)$$

We set $\phi_{follows}(\mathbf{v}_u) = \mathbf{W}_{follows} \cdot \mathbf{v}_u + \mathbf{b}_{follows}$, where $\mathbf{W}_{follows}$ ’s norm $\|\mathbf{W}_{follows}\|$ is limited. Given all the user-user factoids in $\mathcal{F}_{follows}$, i.e. $\{\langle u_i, follows, u_j \rangle\}$, we define the following objective function.

$$f(\mathcal{F}_{follows}) = \sum_{\langle u_i, follows, u_j \rangle \in \mathcal{F}_{follows}} \log(prob(u_i, follows, u_j)) \quad (5)$$

We learn user embedding vectors $\{\mathbf{v}_u\}_{u \in \mathcal{U}^a}$ by solving the following multi-objective optimization problem.

$$\max_{\{\mathbf{v}_u\}} (f(\mathcal{F}_{follows}), f(\mathcal{F}_{pred_1}), \dots, f(\mathcal{F}_{pred_d})) \quad (6)$$

Once we learned $\{\mathbf{v}_u\}_{u \in \mathcal{U}^a}$, we link user identities from different OSNs by simply comparing the distance between their embedding vectors. For example, in the user embedding space in Figure 1, as user ID 1’s nearest neighbor is user ID 6, we link them as the same person.

D. Optimization

To solve the multi-objective optimization in Equation 6, we optimize $f(\mathcal{F}_{follows}), f(\mathcal{F}_{pred_1}), \dots, f(\mathcal{F}_{pred_d})$ in turn.

As optimizing each objective function $f(\mathcal{F})$ is computationally expensive, we adopt the approach of negative sampling proposed in [3]. Particularly, for each factoid $\langle u_i, \cdot, \cdot \rangle$, K “fake” factoids are introduced, i.e. $\{\langle u_k, \cdot, \cdot \rangle\}$, where u_k are sampled from some noise distribution $P(u)$. More specifically, for a user-user factoid $\langle u_i, follows, u_j \rangle \in \mathcal{F}_{follows}$, we specifies the following objective function for it:

$$f(u_i, follows, u_j) = \log \sigma(\mathbf{v}_{u_i}^\top \cdot \phi_{follows}(\mathbf{v}_{u_j})) + \sum_{k=1}^K E_{u_k \sim P_1(u)} [\log \sigma(-\mathbf{v}_{u_k}^\top \cdot \phi_{follows}(\mathbf{v}_{u_j}))] \quad (7)$$

where $\sigma(x) = \frac{1}{1 + \exp(-x)}$ is the sigmoid function. The first term models the observed factoid, and second term models the “fake” factoids and K is the number of “fake” factoids. We set $P_1(u) \propto d_u^{3/4}$ as proposed in [3], where d is the out-degree of u in the unified network. For a user-object factoid $\langle u_i, pred, o \rangle \in \mathcal{F}_{pred}$, its objective function is as follows.

$$f(u_i, pred, o) = \log \sigma(\mathbf{v}_{u_i}^\top \cdot \phi_{pred}(\mathbf{v}_o)) + \sum_{k=1}^K E_{u_k \sim P_2(u)} [\log \sigma(-\mathbf{v}_{u_k}^\top \cdot \phi_{pred}(\mathbf{v}_o))] \quad (8)$$

And we set $P_2(u)$ as a uniform distribution over \mathcal{U}^a .

Then stochastic gradient descent is used to optimize Equations 7 and 8. Algorithm 1 gives an overview of Factoid Embedding. Suppose object embeddings and user embeddings have the same dimension m . The time complexity for each update operation in Algorithm 1 is $O(K \cdot m^2)$. So the time complexity for Algorithm 1 goes through all the user-object factoids and all the user-user factoids once are $O((|\mathcal{U}^s| + |\mathcal{U}^t|) \cdot d \cdot K \cdot m^2)$ and $O((|\mathcal{E}^s| + |\mathcal{E}^t|) \cdot K \cdot m^2)$ respectively.

When we optimize $f(u_i, pred, o)$, we actually push \mathbf{v}_{u_i} in the direction of $\phi_{pred}(\mathbf{v}_o)$. It means that, user identities who share similar objects will be pushed towards each other. (There is also a similar effect for $f(u_i, follows, u_j)$.) This explains why Factoid Embedding is able to push similar user identities close to each other in the user embedding space.

ALGORITHM 1: Factoid Embedding

Input: $\mathcal{F}_{follows}, \mathcal{F}_{pred_1}, \dots, \mathcal{F}_{pred_d}$: the factoids with different predicates.
Input: $\{\mathbf{v}_o\}_{o \in O_{pred_1}}, \dots, \{\mathbf{v}_o\}_{o \in O_{pred_d}}$: the object embeddings for user-object predicates: $pred_1, \dots, pred_d$.
Output: $\{\mathbf{v}_u\}_{u \in \mathcal{U}^a}$.

- 1 Initialize $\{\mathbf{v}_u\}_{u \in \mathcal{U}^a}, \mathbf{W}_{follows}, \mathbf{b}_{follows}, \{\mathbf{W}_{pred_i}\}_{i=1}^d, \{\mathbf{b}_{pred_i}\}_{i=1}^d$;
- 2 **repeat**
- 3 **for** $pred \in \{pred_1, \dots, pred_d\}$ **do**
- 4 sample a batch of user-object factoids \mathcal{F}_{pred}^B from \mathcal{F}_{pred} ;
- 5 **for** $\langle u_i, pred, o \rangle \in \mathcal{F}_{pred}^B$ **do**
- 6 sample K “fake” factoids $\{\langle u_k, pred, o \rangle\}_{k=1}^K$;
- 7 update \mathbf{v}_u according to $\frac{\partial f(u_i, pred, o)}{\partial \mathbf{v}_u}$;
- 8 **end**
- 9 update \mathbf{W}_{pred} and \mathbf{b}_{pred} (once for a certain # of iterations);
- 10 **end**
- 11 sample a batch of user-user factoids $\mathcal{F}_{follows}^B$ from $\mathcal{F}_{follows}$;
- 12 **for** $\langle u_i, follows, u_j \rangle \in \mathcal{F}_{follows}^B$ **do**
- 13 sample K “fake” factoids $\{\langle u_k, follows, u_j \rangle\}_{k=1}^K$;
- 14 update \mathbf{v}_u according to $\frac{\partial f(u_i, follows, u_j)}{\partial \mathbf{v}_u}$;
- 15 **end**
- 16 update $\mathbf{W}_{follows}$ and $\mathbf{b}_{follows}$ (once for a certain # of iterations);
- 17 **until** convergence or reach maximum # of iterations;
- 18 **return** $\{\mathbf{v}_u\}_{u \in \mathcal{U}^a}$.

III. EXPERIMENT

A. Data Collection

We evaluate our proposed Factoid Embedding using data sets from three popular OSNs, namely, Twitter, Facebook and Foursquare. We first gathered a set of Singapore-based Twitter users who declared Singapore as location in their user profiles. From the Singapore-based Twitter users, we retrieve a subset of Twitter users who declared their Facebook or Foursquare accounts in their short bio description as the ground truth. Table II summarizes the statistics of our dataset.

B. Factoid Generation & Object Embedding

The user-user and user-object factoids as described in Section II-A are generated using the user information from the OSNs used in our experiment.

We then calculate the similarity matrices for the data objects. We use Jaro-Winkler distance [2] to measure username and screen name similarity. To measure the similarities between two profile images, we first use the deep learning model VGG16 with weights pre-trained on ImageNet² to

²<https://keras.io/applications/#vgg16>

TABLE II: Datasets.

Dataset	Facebook-Twitter		Foursquare-Twitter	
Network	Facebook	Twitter	Foursquare	Twitter
# Users	17,359	20,024	21,668	25,772
# Links	224,762	165,406	312,740	405,590
Available Information	username,screen name, profile image, network		screen name, profile image, network	
# Ground truth matching pairs	1,998		3,602	

extract a feature vector for each profile image. The cosine similarity between the two profile image feature vectors is then computed. Finally, we embed the data objects $\{\mathbf{v}_o\}_{o \in \mathcal{O}_{pred}}$ for each $pred$ (e.g. username) by using stochastic gradient descent to minimize $error_{pred}$ in Equation 1.

C. Evaluation Baselines and Metrics

1) Supervised Methods.

- **ULink** [4] (S1): a supervised method which models the map from the observed data on the varied social platforms to the latent user space. The node representations learned by Deepwalk³, concatenated with other object embedding vectors are used as user identity features. The code provided by the author is used for UIL.
- **Logistic Regression (LR)** (S2): The following features are used: username similarity, screen name similarity, profile image similarity and the social status in network as defined in [5].

2) Semi-Supervised Methods.

- **COSNET** [5] (SS1): an energy-based model which considers both local and global consistency among multiple networks. The candidate matching graph is generated based on profile-based features: username, screen name and profile image. The public code is used for UIL⁴.
- **IONE** [6] (SS2): a network embedding based approach. Ground truth matching user identity pairs are needed to transfer the context of network structure from the source network to the target network. The original version of IONE uses network information only for UIL. For a fair comparison, we introduce more anchor links by linking user identities which share the same username, screen name or profile image.
- **Factoid Embedding* (FE*)** (SS3): Our proposed Factoid Embedding with labeled matching user identity pairs. Specifically, we adapt our solution to a semi-supervised version by merging the matching user identities into one node in the unified network. The merged user identities therefore share the same embedding vectors.

3) Unsupervised Methods.

- **Name** (U1): an unsupervised approach based on name similarity, which is reported as the most

discriminative feature for UIL [7]. Here it can refer to username or screen name. We present whichever has the better performance.

- **CNL** [8] (U2): An unsupervised method which links users across different social networks by incorporating heterogeneous attributes and social features in a collective manner. The code provided by the author is used for UIL.
- **Factoid Embedding (FE)** (U3): Our proposed Factoid Embedding without any labeled matching user identity pairs.

For each ground truth matching pairs (u^{s*}, u^{t*}) , we rank all the target users, i.e. $u^t \in \mathcal{G}^t$ according to $\cos(\mathbf{v}_{u^{s*}}, \mathbf{v}_{u^t})$. To quantitatively evaluate this ranking, we employ the following two metrics:

- **HitRate@K (HR@K)** in which a ranking is considered as correct if the matching user identity u^{t*} is within the top K candidates, i.e. $rank(u^{t*}) \leq K$.
- **Mean Reciprocal Rank (MRR)** is defined as follows.

$$MRR = \frac{1}{n} \sum_{(u^{s*}, u^{t*})} \frac{1}{rank(u^{t*})}$$

where (u^{s*}, u^{t*}) is a ground truth pair, and n is the number of all the ground truth pairs.

D. Experimental Results

Prediction Performance. We randomly partition the ground truth matching user identity pairs into five groups and conduct five-fold cross-validation. Table III presents the overall performance of the comparison methods on the Facebook-Twitter data set. It shows that, our proposed Factoid Embedding (SS3/U3) yields the best MRR result. Although ULink performs best on HR@1, Factoid Embedding outperforms it on both HR@K and MRR. The reason may be that, as a supervised approach ULink may link precisely the user identity pairs which can be represented by the training dataset. However, for the user identity pairs outside the labeled matching pairs, ULink may lose the ability to match them correctly. In contrast, by embedding factoids, Factoid Embedding is able to link such user identity pairs in an unsupervised manner. It explains why ULink has highest HR@1 but relatively low HR@30. It is a common problem for the supervised solutions for UIL because, as we mentioned in the introduction, the labeled dataset is quite small compared to the whole population. We also can observe that the Factoid Embedding outperforms the existing network embedding approach IONE(SS2), which makes use of the network information only. Interestingly, it

³<https://github.com/phanein/deepwalk>

⁴<https://aminor.org/cosnet>

TABLE III: Performance on Facebook-Twitter Dataset

S/N	Method	HR@1	HR@2	HR@3	HR@4	HR@5	HR@10	HR@30	MRR
S1	ULink	0.7071	0.7285	0.7414	0.7471	0.7557	0.7757	0.8042	0.7102
S2	LR	0.5965	0.6551	0.6906	0.7117	0.7262	0.7837	0.8098	0.6592
SS1	COSNET	0.6586	0.7242	0.7337	0.7367	0.7382	0.7417	0.7452	0.6964
SS2	IONE	0.5605	0.5695	0.5725	0.5730	0.5750	0.5805	0.6031	0.5698
SS3	FE*	0.6851	0.7322	0.7567	0.7747	0.7822	0.8098	0.8508	0.7297
U1	Name	0.5825	0.6226	0.6406	0.6521	0.6626	0.6886	0.7232	0.6201
U2	CNL	0.5930	0.6225	0.6387	0.6451	0.6506	0.6701	0.7327	0.6284
U3	FE	0.6781	0.7292	0.7542	0.7732	0.7827	0.8103	0.8493	0.7254

TABLE IV: Performance on Foursquare-Twitter Dataset

S/N	Method	HR@1	HR@2	HR@3	HR@4	HR@5	HR@10	HR@30	MRR
S1	ULink	0.5464	0.5843	0.6032	0.6232	0.6399	0.6766	0.7397	0.5915
S2	LR	0.5285	0.5913	0.6171	0.6388	0.6473	0.6862	0.7384	0.5882
SS1	COSNET	0.5421	0.5905	0.6116	0.6238	0.6340	0.6585	0.6693	0.5826
SS2	IONE	0.4081	0.4158	0.4225	0.4269	0.4297	0.4408	0.4733	0.4212
SS3	FE*	0.5541	0.6021	0.6293	0.6440	0.6546	0.6979	0.7456	0.6029
U1	Name	0.5227	0.5730	0.5980	0.6154	0.6332	0.6768	0.7293	0.5741
U2	CNL	0.5283	0.5786	0.6050	0.6172	0.6408	0.6877	0.7388	0.5853
U3	FE	0.5433	0.5957	0.6210	0.6374	0.6482	0.6937	0.7423	0.5944

can be seen that the performance of our unsupervised Factoid Embedding (U3) is very close to the semi-supervised version (SS3). One possible explanation is that SS3 just merges the matching pairs into one node, but does not learn from the labeled matching pairs like other supervised solutions e.g. ULink. Realizing that the performance of name similarity (U1) is relatively good, we think, by “pushing” similar user identities close to each other, U3 is able to “merge” these matching pairs in the user embedding space by itself. Thus the performances of U3 and SS3 are not significantly different. Table IV shows the results on the Foursquare-Twitter Dataset, which are consistent to that in Table III. We can therefore conclude that our proposed Factoid Embedding performs best in both the unsupervised and supervised settings.

Parameter Analysis. We investigate the performance w.r.t. the embedding dimension and the number of iterations on the Facebook-Twitter dataset. Figure 2(c) shows that the MRR performance of Factoid Embedding improves as the dimension increases. Figure 2 (d) shows the MRR performances over different numbers of iterations. We can see that the performance improves more significantly in the early iterations.

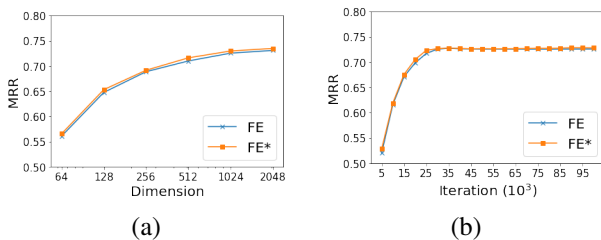


Fig. 2: (a) Performance over Dimensions of User Embedding. (b) Performance over Iterations.

IV. CONCLUSION

In this paper, we proposed a novel framework Factoid Embedding, which adopts an unsupervised approach to cope with heterogeneity in user information and link users identities across multiple OSNs. We evaluated Factoid Embedding using real-world datasets from three OSNs and benchmarked against the state-of-the-art UIL solutions. Our experimental results show that Factoid Embedding outperforms the state-of-the-art UIL solutions even in situations where the names of the user identities are dissimilar.

ACKNOWLEDGMENT

This research is supported by the National Research Foundation, Prime Minister’s Office, Singapore under its International Research Centres in Singapore Funding Initiative.

REFERENCES

- [1] K. Shu, S. Wang, J. Tang, R. Zafarani, and H. Liu, “User identity linkage across online social networks: A review,” *SIGKDD Explorations*, vol. 18, no. 2, 2016.
- [2] W. W. Cohen, P. Ravikumar, and S. E. Fienberg, “A comparison of string distance metrics for name-matching tasks,” in *IJCAI*, 2003.
- [3] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *NIPS*, 2013.
- [4] X. Mu, F. Zhu, E. Lim, J. Xiao, J. Wang, and Z. Zhou, “User identity linkage by latent user space modelling,” in *SIGKDD*, 2016.
- [5] Y. Zhang, J. Tang, Z. Yang, J. Pei, and P. S. Yu, “COSNET: connecting heterogeneous social networks with local and global consistency,” in *SIGKDD*, 2015.
- [6] L. Liu, W. K. Cheung, X. Li, and L. Liao, “Aligning users across social networks using network embedding,” in *IJCAI*, 2016.
- [7] A. Malhotra, L. C. Totti, W. M. Jr., P. Kumaraguru, and V. A. F. Almeida, “Studying user footprints in different online social networks,” in *ASONAM*, 2012.
- [8] M. Gao, E. Lim, D. Lo, F. Zhu, P. K. Prasetyo, and A. Zhou, “CNL: collective network linkage across heterogeneous social platforms,” in *ICDM*, 2015.