

Locate Who You Are: Matching Geo-location to Text for User Identity Linkage

Jiangli Shao^{1,2}, Yongqing Wang¹, Hao Gao^{1,2}, Huawei Shen¹, Yangyang Li³, Xueqi Cheng¹

¹Data Intelligence System Research Center, Institute of Computing Technology, Chinese Academy of Sciences

²University of Chinese Academy of Sciences

³National Engineering Lab for Risk Perception and Prevention, China Academic Electronic and Information Technology
{shaojiangli19z,wangyongqing,gaohao,shenhuawei,cxq}@ict.ac.cn,liyangyang@cetc.com.cn

ABSTRACT

Nowadays, users are encouraged to activate across multiple online social networks simultaneously. User identity linkage, which aims to reveal the correspondence among different accounts across networks, has been regarded as a fundamental problem for user profiling, marketing, cybersecurity, and recommendation. Existing methods mainly address the prediction problem by utilizing profile, content, or structural features of users in symmetric ways. However, encouraged by online services, information from different social platforms may also be asymmetric, such as geo-locations and texts. It leads to an emerged challenge in aligning users with asymmetric information across networks. Instead of similarity evaluation applied in previous works, we formalize correlation between geo-locations and texts and propose a novel user identity linkage framework for matching users across networks. Moreover, our model can alleviate the label scarcity problem by introducing external text-location pairs. Experimental results on real-world datasets show that our approach outperforms existing methods and achieves state-of-the-art results.

CCS CONCEPTS

• **Information systems** → *Location based services; Data mining.*

KEYWORDS

user identity linkage; geo-location; user generated text

ACM Reference Format:

Jiangli Shao, Yongqing Wang, Hao Gao, Huawei Shen, Yangyang Li, Xueqi Cheng. 2021. Locate Who You Are: Matching Geo-location to Text for User Identity Linkage. In *Proceedings of the 30th ACM Int'l Conf. on Information and Knowledge Management (CIKM '21), November 1–5, 2021, Virtual Event, Australia*. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3459637.3482134>

1 INTRODUCTION

Nowadays, the highly developed online techniques make progress in social media applications, encouraging users to share diverse information on different social platforms. User identity linkage

(UIL), which aims to reveal the correspondence among different accounts across networks, has been in the spotlight for a long time due to the important role in several downstream tasks, such as user profiling[18], marketing[17], cybersecurity[13] and recommendation [2].

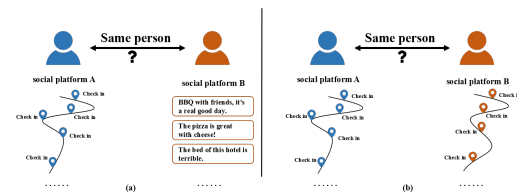


Figure 1: A sketch to show the differences between UIL tasks with asymmetric information(a) and symmetric information(b).

Traditionally, most existing approaches in the literature attempt to address the linkage problem with kinds of information, such as profile, content, and structural data[4, 13]. For instance, researchers extract delicate features from user name, age, gender, registration time, location, and email, and then construct a well-defined prediction model across networks[5]. Moreover, Nie et al.[10] identified correspondence by learning latent interests from user-generated content. Recently, advances in network embedding techniques drive the progress of UIL modeling by structural features. Some more powerful models have emerged such as PALE[8], DeepLink[21], and PAAE[12]. Furthermore, by making use of multi-information, LHNE[16] and TALP[7] attempt to model UIL on heterogeneous networks. Although kinds of information are applied in different modeling, the input data across networks must be symmetric, that is, only those users who have similar types of information in profile, content, and structures can be aligned across networks.

However, users are also encouraged to post different types of information on different social platforms. For example, a user can travel a scenic spot, post a tweet on Twitter and check in on Foursquare. The scenario makes a new challenge for UIL in practice, that is, how to align user identification with so-called "asymmetric information" across networks. Due to the discrepancy of feature spaces, traditional methods may perform worse or even become invalid in this asymmetric situation.

In this paper, we are interested in a common UIL problem based on asymmetric information, i.e., *can we align user identification with texts and geo-locations across network?* To solve the problem, two key issues should be approached ahead: 1) how to evaluate the correlation between texts and geo-locations? 2) how to match pair of

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CIKM '21, November 1–5, 2021, Virtual Event, Australia.

© 2021 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-8446-9/21/11...\$15.00

<https://doi.org/10.1145/3459637.3482134>

users according to sequences generated by texts and geo-locations? Initially, the text-location correlation matrix can be calculated by the co-occurrence between geo-location and corresponding text expression across networks. However, such co-occurrence is hard to be detected especially in the situation with scarce labeled linkage between users. Therefore, external data referring to text-location relationship is also applied in modeling text-location correlation matrix. Based on text-location correlation matrix, during matching stage, a user-user interactive tensor is constructed with user pairs from text sequence to geo-location sequence. On the basis of interactive tensor, a 3D convolution neural network is used to determine whether the pair of two accounts is linked or not. In two real applications, the prediction accuracy resulted by the proposed model is over 89%, outperforming the state-of-the-art methods modeled by symmetric information. With supplementary of additional text-location relationship from external data, interestingly, the experimental results show a significant improvement, especially when training on datasets with scarce linkage labels. Thus, it shows great potentials in real applications. In a nutshell, the main contributions of our work can be listed as follows:

- A new challenge on utilizing asymmetric information for UIL is defined. In the scenario of aligning users with texts and geo-locations across networks, a novel UIL model is proposed by associating text-location information. To our knowledge, this is the first work to utilize asymmetric information in UIL tasks.
- To circumvent linkage label sparsity, external data referring to text-location relationship is applied in modeling, leading at most 11% improvement in datasets with scarce linkage labels.
- Interestingly, experiments on two real-world dataset show the experimental results from models with asymmetric information outperforms the results from state-of-the-art models with symmetric information, proving great potentials in real applications.

2 RELATED WORK

Previous works on UIL with heterogeneous information are briefly introduced in this section. With heterogeneous information, Zhang et al.[20] formalize a heterogeneous network and implement random walk to explore same user identities. Recently, the advances in network embedding promote the analysis of heterogeneous information. LHNE[16] embeds cross-network structural and content information into a unified space by jointly capturing the friend-based and interest-based user co-occurrence in intra-network and inter-network, respectively. Then the learned embedding vectors can be applied in matching user identities. With trajectory data, DPLink[3] solves UIL problem based on heterogeneous mobility data collected from services. However, heterogeneous information applied in those mentioned models still requires to be used symmetrically across networks. Thus, the matching performance is still worse in asymmetric scenario.

Meanwhile, the idea of interactive tensor is also applied in natural language processing [11, 15]. Nevertheless, the proposed interactive tensor, in this paper, concentrates more on exploring correlation between texts and geo-locations. Moreover, the built tensor is

sensitive in time, promising better prediction performance in real applications.

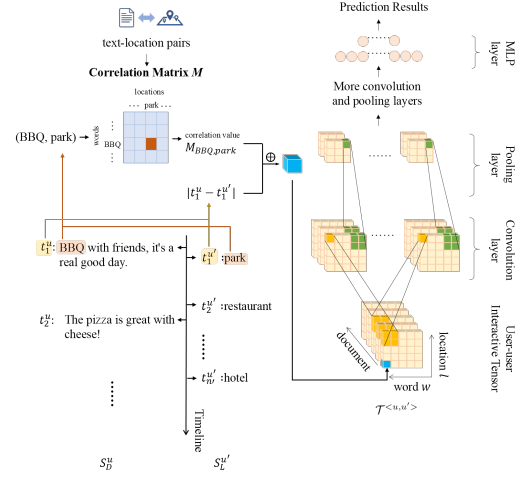


Figure 2: The framework of our model, which takes user record sequences S_D^u and $S_L^{u'}$ as inputs and outputs the probability that user u and u' belong to the same natural person.

3 METHODOLOGY

In this section, problem statement is introduced at first. Then, the specifics of proposed model are presented.

3.1 Problem Formulation

In the scenario of aligning users with texts and geo-locations across networks, users can check in at spots on one online service L and post corresponding text messages on another online service D . Generally, the trajectory of user u can be recorded as chronological sequence $S_L^u = \{(l_k^u, t_k^u) | k = 1 \dots N_L\}$, where l_k^u and t_k^u in tuple (l_k^u, t_k^u) refer to the k -th location and time generated by user u . Meanwhile, posts of user u can be represented as another chronological sequence $S_D^u = \{(d_k^u, t_k^u) | k = 1 \dots N_D\}$, where d_k^u and t_k^u refer to the k -th document and its publish time generated by user u , and document d_k is consisted of word sequence $w_{k1} \dots w_{km}$.

For each user identity u in one platform (e.g., serving texting services), the objective of UIL with asymmetric information is finding a mapping function f to identify, if any, its counterpart u' in another platform (e.g., serving location services). The mapping function f can be formalized as,

$$f(S_D^u, S_L^{u'}) = \begin{cases} 1, & \text{if } u \text{ and } u' \text{ belong to the same person,} \\ 0, & \text{otherwise.} \end{cases}$$

3.2 Model Framework

Before diving into the detail of the proposed model, the difference of modeling symmetric and asymmetric information is illustrated firstly. When tackling UIL tasks with symmetric information, similarity can be conveniently evaluated by kinds of metrics defined on

distances. However, the pairs of asymmetric data can hardly be evaluated by those distance metrics. Instead of distance metrics, a text-location correlation matrix M between words and geo-locations is calculated according to text-location pairs. Given labeled pairs of texts and geo-locations, the correlation value of word w and location l can be defined in a *tf-idf* way, such as:

$$M_{w,l} = \log(1 + \text{count}(w, l)) \cdot \log \left(1 + \frac{\sum_{l' \in L} \text{count}(w, l')}{\epsilon + \text{count}(w, l)} \right). \quad (1)$$

Here, $\text{count}(w, l)$ denotes the number of occurrences between word w and location l appeared in all documents and their associated location. The ϵ is a constant.

Based on the calculated correlation matrix M , a user-user interactive tensor $\mathcal{T}^{<u, u'>}$ can be defined as inputs of S_D^u and $S_L^{u'}$. Formally, every element in $\mathcal{T}^{<u, u'>}$ is formulated as:

$$\mathcal{T}_{ijk}^{<u, u'>} = M_{w_{ik}, l_j} \oplus |t_i^u - t_j^{u'}|. \quad (2)$$

$$1 \leq i \leq |S_D^u|, 1 \leq j \leq |S_L^{u'}|, 1 \leq k \leq K, i, j, k \in \mathbb{Z}^+.$$

Here, \oplus represents the concatenation operation and operator $|x|$ means the absolute value of number x . Note that the length of document is truncated or padded to K in order to implement the following convolution operation.

The 3D convolutional operation[9] and dynamic pooling[14] strategy are then used to extract different levels of matching patterns. Let z^l denote the feature map in l -th layer and $z^0 = \mathcal{T}^{<u, u'>}$. Convolution layers and dynamic pooling layers are defined as follows:

$$C_{i,j,k}^{(l,n)} = \sigma \left(\sum_{n'=0}^{N_k-1} \sum_{p=0}^{K_p-1} \sum_{q=0}^{K_q-1} \sum_{r=0}^{K_r-1} w_{p,q,r}^{(l,n)} \cdot z_{i+p,j+q,k+r}^{(l-1,n')} + b^{(l,n)} \right), \quad (3)$$

$$z_{i,j,k}^{(l,n)} = \max_{0 \leq p < d_l} \max_{0 \leq q < d_w} \max_{0 \leq r < d_h} C_{i \cdot d_l + p, j \cdot d_w + q, k \cdot d_h + r}^{(l,n)}. \quad (4)$$

Here, $C^{(l,n)}$ is the convolutional output of n -th kernels in l -th layers. The $\sigma(\cdot)$ denotes ReLU activation function. The size of the kernel is $K_p * K_q * K_r$ and N_k is the number of channels. Meanwhile, the length, width and height of pooling operator at l -th layer is denoted as d_l, d_w and d_h , determined by the size of l -th convolutional feature map $\alpha * \beta * \gamma$ and the size of pooled feature map $\alpha' * \beta' * \gamma'$, i.e., $d_l = \lceil \alpha / \alpha' \rceil, d_w = \lceil \beta / \beta' \rceil, d_h = \lceil \gamma / \gamma' \rceil$.

Eventually, We use a MLP (Multi-Layer Perception) to produce the final prediction results:

$$\hat{y} = \text{sigmoid}(\mathbf{W} \cdot \mathbf{z} + \mathbf{b}). \quad (5)$$

During training process, the objective function is formulated by cross entropy, i.e., minimizing the loss

$$\text{Loss} = - \sum_{i=1}^N [y^{(i)} \ln(\hat{y}^{(i)}) + (1 - y^{(i)}) \ln(1 - \hat{y}^{(i)})], \quad (6)$$

where $y^{(i)}$ is the label of the i -th training instance and N is the number of training samples.

Overall, the framework of the proposed model is briefly depicted in Figure 2.

4 EXPERIMENT

Datasets The experiments are conducted on two real-world datasets. The first one is the Twitter-Foursquare dataset, which is collected by [19]. This dataset contains the trajectories data from Foursquare, a popular location-based social network, and Twitter, a micro-blog social network around the world. Null values are left out in this dataset and there remain 2044 linked user accounts in total. The other dataset comes from Dazhong Dianping¹, a popular Chinese social application with over ten million registered users. To build this dataset, we collect 2218 users' review and check-in messages at Dazhong from the year 2012 to 2014 and regard them as two types of input information. Moreover, review and check-in data published by same user at same timestamp are discarded to imitate the behavior of users on two distinct platforms in real world. More details about the two datasets are listed in Table 1.

Table 1: A brief description of two datasets

	#users	#records		#users	#records
Foursquare	2044	37061	Dazhong check-in	2218	82605
Twitter	2044	4084874	Dazhong review	2218	74821

Baselines We compare our models with the following four baselines. Network structure features are not used in all of the models to ensure the fairness of comparison and hyper-parameters of baselines are the same as those in the original paper.

MNA[6]: This model extracts text features, temporal features, and spatial features from input sequences and sends them to an SVM to finally give the predicted results.

STUL[1]: STUL captures stay regions and corresponding time distribution features from input sequences and identifies linked user accounts by comparing similarities in these two features.

DPLINK and DPLINK*[3]: DPLINK is an end-to-end deep neural network model that uses LSTM to capture features in the input sequence and use co-attention mechanisms to fusion two trajectories. DPLINK* is a variant of the original DPLINK model in which different types of features are utilized as the input of LSTM.

Settings In the experiments, the text data is cleaned by removing stopwords, punctuation, and URLs. For the Twitter-Foursquare dataset, the length of tweets is limited to 50 and each Foursquare geo-location is converted to a one-hot vector with 476 dimensions according to its category. For the Checkin-Review dataset, the review length is restricted to 100 after word segmentation and each check-in location is converted to a one-hot vector with 45 dimensions according to its category.

The negative sample is constructed by choosing non-linked users from two platforms. The number of positive and negative samples keeps the same. We use 80% data for training, 10% for validating, and 10% for testing. As for the hyperparameter settings, 2 convolutional layers and 2 pooling layers are used in model for both datasets. Kernel size is set to $3*3*3$ and the number of channels is 8. Performances of models are evaluated by F1, accuracy(ACC), and AUC metrics.

Results Firstly, we make an overall comparison of the prediction performance among all of these models. The results are shown in Table 2. We can notice that our model outperforms the other

¹<https://www.dianping.com/>

Table 2: Performance comparison of different methods. The "Type" column shows the type of information used in UIL tasks in which "sym" denotes symmetric information and "asym" denotes asymmetric information.

	Type	Twitter-Foursquare			Checkin-Review		
		F1	ACC	AUC	F1	ACC	AUC
MNA	sym	0.7278	0.7415	0.7720	0.8917	0.8919	0.9573
STUL	sym	0.7563	0.7683	0.8573	0.9099	0.9099	0.9605
DPLINK	sym	0.8242	0.8244	0.8838	0.9459	0.9459	0.9802
DPLINK*	asym	0.7826	0.7854	0.8788	0.9030	0.9032	0.9716
OURS	asym	0.8926	0.8927	0.9327	0.9685	0.9684	0.9892

four baselines on both of the two datasets. The proposed model obtains absolute improvements for about 6.8% and 2.3% in terms of F1, 6.8% and 2.3% in terms of ACC, and 4.9% and 0.1% in terms of AUC when compared with the best baseline on two datasets respectively, demonstrating the effectiveness of the proposed framework in UIL task. According to results from MNA, STUL, DPLINK and our model, traditional models with symmetric information have a poor performance in this asymmetric situation as discrepancies between the text feature space and location feature space. However, our model avoids the similarity comparison between two different feature spaces via computing correlation between words and geo-locations, which might explain the excellent results of our model on this task. Moreover, the proposed model explores more precise matching patterns in word-location level, thus enabling better performance than DPLINK*, which extracts matching patterns in sequence representation level.

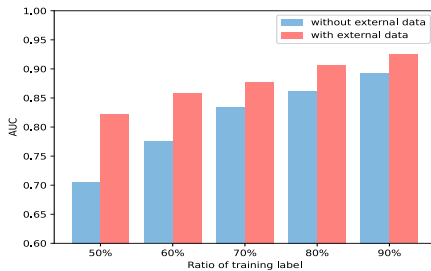


Figure 3: Effects of adding external data to our model over different ratios of training label

According to the correlation matrix defined in Eq. 1, the proposed correlation matrix can also be calculated by relevant text-location pairs. As a consequence, we introduce the yelp² review data and foursquare review data for better estimating the correlation matrix. With the additional text-location pairs, the estimation bias caused by scarce linkage data can be reduced to some extent. As depicted in Figure 3, with the supplementary data, the performance on AUC is increased by 11.7%, 8.4%, 4.3%, 4.5% and 3.2% when using 50%, 60%, 70%, 80% and 90% training label respectively. Interestingly, according to the improvement on different settings, the introduced data can take more positive effects on scarce linkage labels. It is practical to be applied in real application.

²A crowd-sourced review site for businesses. Website: <https://www.yelp.com>

Finally, to give a deeper insight into the linkage mechanism of the model, we exhibit a case of explored linked user accounts in test set in figure 4. The correlation between words and locations is highlighted in the figure. We can observe that, when visiting some places, users are likely to check in on one platform and post some related tweets on the other at a similar time. In other words, during some periods, records on two platforms show not only temporal but also semantic correlation. The findings are consistent with our intuition and serve as the fundamental matching patterns of the proposed model. By detecting and aggregating such kinds of matching patterns, the framework can correctly discover linked user accounts.

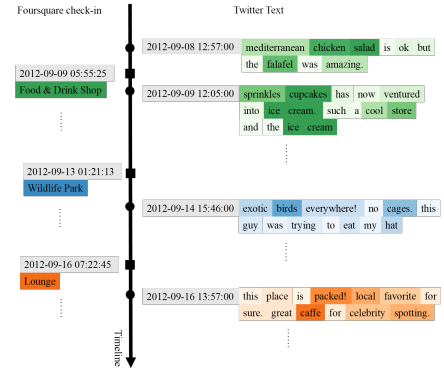


Figure 4: A case to show the results of the proposed model. Squares and circles on the timeline represent the check-in on foursquare and tweets on twitter respectively. Different colors of foursquare check-in mark different locations and the corresponding color of words in tweets indicate the value in correlation matrix. The darker the color, the higher the correlation between words and locations.

5 CONCLUSION

In this paper, we propose a novel user identity linkage model by matching geo-locations and texts in two social platforms. With inputs of asymmetric information, the proposed model will first construct a user-user interactive tensor via times and the correlation matrix between words and locations. Then, 3D convolutional operations are conducted on the tensor to capture the distinct features. Experimental results show that our model can greatly improve the user identity linkage performance compared with baselines. Meanwhile, we observe that the proposed framework can overcome the training linkage label scarcity problem to some extent by adding additional text-location pairs to calculate the correlation matrix of words and geo-locations, showing great potentials in real applications.

ACKNOWLEDGMENTS

This work was funded by the National Natural Science Foundation of China under grant numbers 61802371, 91746301, U1836111, the National Key Research and Development Program of China under grant numbers 2018YFC0825200 and the National Social Science Fund of China under grant number 19ZDA329.

REFERENCES

- [1] Wei Chen, Hongzhi Yin, Weiqing Wang, Lei Zhao, Wen Hua, and Xiaofang Zhou. 2017. Exploiting spatio-temporal user behaviors for user linkage. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 517–526.
- [2] Shaohua Fan, Junxiong Zhu, Xiaotian Han, Chuan Shi, Linmei Hu, Biyu Ma, and Yongliang Li. 2019. Metapath-guided heterogeneous graph neural network for intent recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2478–2486.
- [3] Jie Feng, Mingyang Zhang, Huandong Wang, Zeyu Yang, Chao Zhang, Yong Li, and Depeng Jin. 2019. Dplink: User identity linkage via deep neural network from heterogeneous mobility data. In *The World Wide Web Conference*. 459–469.
- [4] Hao Gao, Yongqing Wang, Shanshan Lyu, Huawei Shen, and Xueqi Cheng. 2020. GCN-ALP: Addressing Matching Collisions in Anchor Link Prediction. In *2020 IEEE International Conference on Knowledge Graph, ICKG 2020, Online, August 9–11, 2020*, Enhong Chen and Grigoris Antoniou (Eds.). IEEE, 412–419. <https://doi.org/10.1109/ICKG50248.2020.00065>
- [5] Oana Goga, Howard Lei, Sree Hari Krishnan Parthasarathi, Gerald Friedland, Robin Sommer, and Renata Teixeira. 2013. Exploiting innocuous activity for correlating users across sites. In *Proceedings of the 22nd international conference on World Wide Web*. 447–458.
- [6] Xiangnan Kong, Jiawei Zhang, and Philip S Yu. 2013. Inferring anchor links across multiple heterogeneous social networks. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. 179–188.
- [7] Xiaoxue Li, Yanmin Shang, Yanan Cao, Yangxi Li, Jianlong Tan, and Yanbing Liu. 2020. Type-Aware Anchor Link Prediction across Heterogeneous Networks Based on Graph Attention Network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 147–155.
- [8] Tong Man, Huawei Shen, Shenghua Liu, Xiaolong Jin, and Xueqi Cheng. 2016. Predict anchor links across social networks via an embedding approach. In *Ijcai*, Vol. 16. 1823–1829.
- [9] Pavlo Molchanov, Shalini Gupta, Kihwan Kim, and Jan Kautz. 2015. Hand gesture recognition with 3D convolutional neural networks. In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 1–7. <https://doi.org/10.1109/CVPRW.2015.7301342>
- [10] Yuanping Nie, Yan Jia, Shudong Li, Xiang Zhu, Aiping Li, and Bin Zhou. 2016. Identifying users across social networks based on dynamic core interests. *Neurocomputing* 210 (2016), 107–115.
- [11] Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, Shengxian Wan, and Xueqi Cheng. 2016. Text matching as image recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30.
- [12] Yanmin Shang, Zhezhou Kang, Yanan Cao, Dongjie Zhang, Yangxi Li, Yang Li, and Yanbing Liu. 2019. PAEE: A unified framework for predicting anchor links with adversarial embedding. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 682–687.
- [13] Kai Shu, Suhang Wang, Jiliang Tang, Reza Zafarani, and Huan Liu. 2017. User identity linkage across online social networks: A review. *Acm Sigkdd Explorations Newsletter* 18, 2 (2017), 5–17.
- [14] Richard Socher, Eric H Huang, Jeffrey Pennington, Andrew Y Ng, and Christopher D Manning. 2011. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *NIPS*, Vol. 24. 801–809.
- [15] Shengxian Wan, Yanyan Lan, Jiafeng Guo, Jun Xu, Liang Pang, and Xueqi Cheng. 2016. A deep architecture for semantic matching with multiple positional sentence representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30.
- [16] Yaqing Wang, Chunyan Feng, Ling Chen, Hongzhi Yin, Caili Guo, and Yunfei Chu. 2019. User identity linkage across social networks via linked heterogeneous network embedding. *World Wide Web* 22, 6 (2019), 2611–2632.
- [17] Yongqing Wang, Huawei Shen, Jinhua Gao, and Xueqi Cheng. 2019. Learning Binary Hash Codes for Fast Anchor Link Retrieval across Networks. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13–17, 2019*, Ling Liu, Ryan W. White, Amin Mantrach, Fabrizio Silvestri, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia (Eds.). ACM, 3335–3341. <https://doi.org/10.1145/3308558.3313430>
- [18] Qianyi Zhan, Jiawei Zhang, Philip Yu, and Junyuan Xie. 2017. Community detection for emerging social networks. *World Wide Web* 20, 6 (2017), 1409–1441.
- [19] Jiawei Zhang, Xiangnan Kong, and Philip S Yu. 2014. Transferring heterogeneous links across location-based social networks. In *Proceedings of the 7th ACM international conference on Web search and data mining*. 303–312.
- [20] Jiawei Zhang, Philip S Yu, and Zhi-Hua Zhou. 2014. Meta-path based multi-network collective link prediction. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1286–1295.
- [21] Fan Zhou, Lei Liu, Kunpeng Zhang, Goce Trajcevski, Jin Wu, and Ting Zhong. 2018. Deeplink: A deep learning approach for user identity linkage. In *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, 1313–1321.