Mapping Users across Social Media Platforms by Integrating Text and Structure Information

Song Sun*†, Qiudan Li*, Peng Yan*, Daniel D.Zeng*†‡
*The State Key Laboratory of Management and Control for Complex Systems
Institute of Automation, Chinese Academy of Sciences, Bejing, China
†University of Chinese Academy of Sciences, Beijing, China
‡Department of Management Information Systems, University of Arizona, Tucson, AZ 85721, USA
Email: {sunsong2015, qiudan.li, yanpeng2017, dajun.zeng}@ia.ac.cn

***Abstract*—With the development of social media technology, users often register accounts, post messages and create friend links on several different platforms. Performing user identity mapping on multi-platform based on the behavior patterns of users is considerable for network supervision and personalization service. The existing methods focus on utilizing either text information or structure information alone. However, text information and structure information reflect different aspects of a user. An organic combination of them is beneficial to mining user behavior patterns, thus help identify users across platforms accurately. The challenging problems are the effective representation and similarity computation of the text and structure information. We propose a mapping method which integrates text and structure information. At first, the model represents user name, description, location information based on word2vec or string matching, and friend information represented as relation network is regarded as structure information. Then these information are used for similarity computation using Jaccard index or cosine similarity. After similarity computation, a linear model is adopted to get the overall similarity of user pairs to perform user mapping. Based on the proposed method, we develop a prototype system, which allows users to set and adjust the weights of different information, or set expected index. The experimental results on a real-world dataset demonstrate the efficiency of the proposed model.**

***Keywords*—user mapping, cross-platform, similarity computation, word2vec.**

I. Introduction

With the prosperity of online social media and social network, people usually tend to have several different accounts on various social network platforms for different reasons [1], such as personal privacy, personal preference and varied advantages of multiple platforms. Mapping users has a significant research value, such as personalized recommendation, advertising recommendation, community detection and discovering. Moreover, it's of great importance to network supervision and regulation, which is a key point to network security. With the help of user mapping, network regulators could do better in controlling the rumors, monitoring the public opinion and tracing the source of news.

Generally speaking, on different online social platforms, human tends to provide same or similar text information such as description, location, username, sex, etc., and according to social network theory, one usually has semblable and overlapping information network including friend, follower and retweet network. Both text information and structure information reflect users behavior pattern. However, most of the existing researches use only text information like user resume [2], username [3] or only structure information like relation network [4] to perform user mapping. In this paper, we aim to perform cross-platform user mapping by obtaining a more comprehensive users behavior pattern based on both text and structure information. The challenging problems are the effective representation and similarity computation of the text and structure information.

In addition to string matching, word2vec [5] is an effective low dimensional vector representation method, thus both string matching and word2vec are adopted to represent the text information such as username, description, and location. Popular similarity measures including Jaccard index and cosine similarity are used to mine latent similarity relationship among users on different platforms. A linear model is designed to obtain the overall similarity of user pairs to perform user mapping.

In summary, the major contributions of this work are: 1) This work is a first step towards integrating both text information and structure information to perform user mapping across social platforms, which can gain better insights into users behavior pattern and thus help identify users accurately. 2) We design an effective representation method for text and structure information based on word2vec and string matching, then develop an integrated linear model to calculate the similarity of the representation and perform user mapping across social platforms. Experimental results on a real-world dataset show the efficiency of the proposed model.

The remainder of this paper is organized as follows. We give a brief summary of related issues and works in Section II. In Section III, we introduce the basic structure of our user mapping framework and discuss the theoretical basis of this framework. And we show the details of the model and how to model the information of users. Next, we show the experimental results of our proposed method in cross-platform user mapping in Section IV. The evaluation results are presented to demonstrate the effectiveness of the method. And in Section V, we conclude this study and give out our promising perspectives regarding future research.

978-1-5090-6727-5/17/$31.00 ©2017 IEEE

II. Related works

User mapping is a newly developing research topic. Existing works mainly include text information based method and structure information based method. The text based method focuses on mining user patterns in text information such as username, location and so on. And structure based method uses friend links and other kinds of relation network information. However, these existing methods rarely use text and structure information together to map users.

A. Text information based mapping method

Text information based method mainly uses text information, like username, location, age, tags and information produced when a user using the online social platforms, like retweets and comments. In [6], the authors introduced a user-mapping approach based on username information called MOBIUS. By using supervised learning, this method establishes a pattern model to study the username pattern in social media. Other text information like tags is used to perform the user mapping, too [7].

Apart from using limited information, some researchers utilize full user record information to map the user across online social platforms. Based on user resume [2], the researchers embed the resume information into vector data and make use of the vector data to compare similar users. This algorithm uses information such as username, location, telephone number, birth date and so on, however, this kind of information is difficult to obtain due to privacy concerns. To address this issue, Liu [3] uses interaction data such as tweets, comments, rating information.

B. Structure information based mapping method

Structure information based mapping method regards users as vertices, and regards friend, retweet, follow relations as edges in the social network graph. Then researchers use the structural properties of the network graph to perform the user mapping. It uses only topological information and thus can protect the user privacy. This kind of methods often regards the user mapping issue as anchor link prediction. An anchor link is a person who uses different online social platforms linked across platforms like an anchor.

There are supervised and unsupervised methods to utilize the relation network. The unsupervised method is a kind of NP-hard combination optimization issue [8]. These methods using only artificially extracted user network features are not efficient. The supervised methods need to know a part of existing anchor links as a training dataset to train the model and use the model to find out the undiscovered anchor links in the network. Some methods directly use artificially extracted features such as degree, clustering coefficient, triangular number and common neighbors [9]. But Man T. [10] et al. think that the methods using such features have not used the internal structure regularity, and its too sensitive to the small changes of the network will lead to strikingly different results. They propose a new method called PALE to learn internal structure regularity of the network and thus make the result more stable. Liu L. [4] et al. use representation learning utilizing several social network information. And they use directed graph rather than common used undirected graph. Tan [11] et al. use an algorithm called MAH to do the anchor link discovering. Zhang Y [12] et al. use COSNET method to do the diversity user network prediction and analysis. On the other hand, Zhang J. [13] et al. use probability network to predict anchor links.

The existing research works focus on either text information or structure information of users, but few uses all confluent information. Our method makes a good use of entire user behavior information and proposes a new method to get the work done.

III. The proposed mapping method

The method we proposed has a theoretical background based on cognitive science and social network theory. Firstly, we effectively represent text information and structure information. For example, user description information is embedded to vectors using word2vec. Then we compute the similarity of all kinds of information between different users using Jaccard index, cosine similarity calculating and string matching. And we use integrated information similarity to accomplish the user mapping tasks.The overall diagram is shown in Fig.1.

A. Theoretical Background

The text information mainly consists of subjective description like user description and objective information like location information. And the structure information mainly consists of social relationship of online social platform users. Both of the information contain useful elements which can be used to analyze the user behavior patterns and compute the similarity of users. And getting a full use of them is the key for accurately mapping users across social platforms.

1) Subjective description: The subjective description is written by the user. In general, there are several forms: a few words of tags, a relatively short section of self-introduction, or a full resume, including a detailed description of the interest or career. Although the form is diverse, the similarity of the content is still guaranteed. This is because the description on the web can basically be an outline of a personal life, which is stable and not easy to change, due to the human's cognitive limitation. As the old saying goes, old habits die hard. When writing online, people tend to complete the task in a short time, so the limitations of dealing with information make description impossible to involve too many aspects of life or full of details. Instead, the description is mainly about user's profession, major or core interests, or other characteristics that are difficult to change.

To sum up, an outline of a personal life derived from the description on social networks can be used for identification.

2) Objective Information: Ultimately, the objective information includes real name, geographic location, country and other factual information. Due to the clarity and simplicity of such information, simple information matching method is

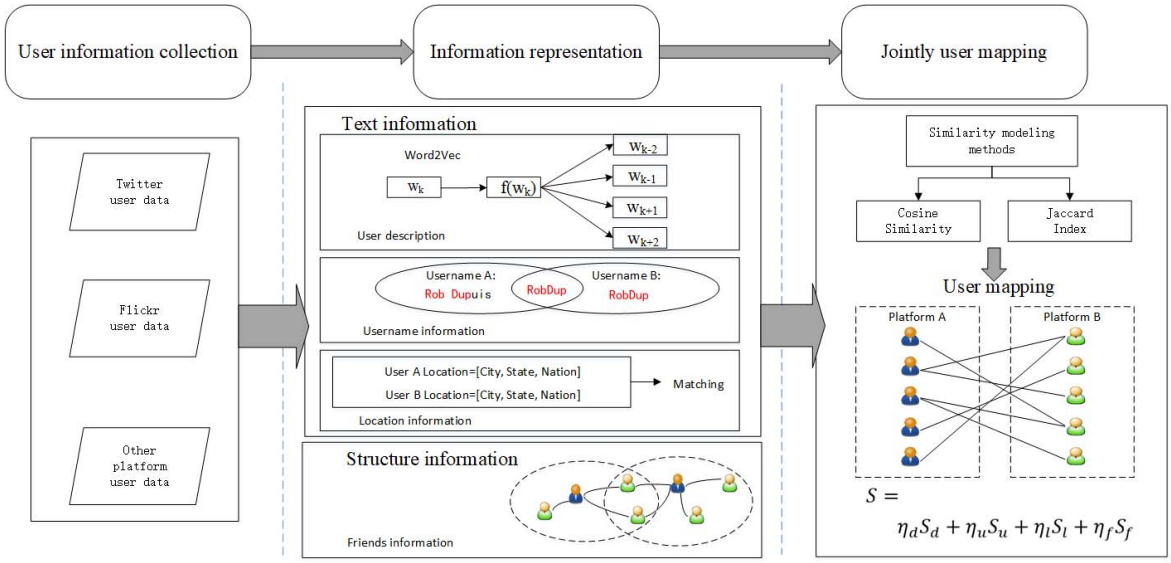

Fig. 1. An overview of the research framework.

appropriate. However, in reality, we find that some people deliberately provide false information to prevent their true identity from being discovered. Thus, on the one hand, we need to assess the credibility of the information, on the other hand, it reflects the necessity to study other identification methods.

3) Social Relationship: The social relationship, one of the most important social characteristics, often represents people's status and role in society, containing rich identification clues. Online social networks record the process of users interacting with others. A graph of social links among users can be built based on this recording information, including friendship or follower-ship/follow-ship. According to the social network theory, each node in the graph model represents a user in the social network, and the lines around nodes represent the relationship between the users. Graph algorithms can be used in the analysis of social network.

Through the construction of the model, the user's social relations and role characteristics are transformed into structural features. For example, the distance of the relationship between two people in society is abstracted as the span between two nodes. An embedding method can use both micro and macro structural regularity for identification.

B. Information representation

1) User description: On most online social network platforms, users can write a short paragraph of description about themselves. The content could be the hobby, the job, or just the full resume. Obviously, the text information of user description counts for much in user similarity measurement. To model the user description information, we use the word2vec model and convert the description text into a vector.

Word2vec is a group of related models that are used to produce word embeddings. By training a large corpus of text, it reconstructs the text information into a vector space, typically of several hundred dimensions. A pair of words having semblable meaning in the corpus will be located in close proximity to one another in the space. Using word2vec, we could solve some more complicated cases, like user description. The descriptions of a user on different platforms are often not the same but have similar text information, and using word2vec could help us extract the information we need and calculate the cosine similarity of the vectors.

After stopwords removal and other language processing, the description of a user will be transferred to a set of words $(word_1, word_2, ..., word_k)$. Then we find the word vector one by one in the pre-trained n-dimension word2vec model, and afterward, we could get a set of embedding n-dimension word vectors $(wordvector_1, wordvector_2, ..., wordvector_k)$.

For each word vector:

$$wordvector = \{v_1, v_2, ..., v_n\} \tag{1}$$

The description of user N on platform A is represented as D_N^A, and it is a n-dimension vector:

$$D_N^A = \{w_1, w_2, ..., w_n\} = \sum_{i=1}^{k} \frac{wordvector_i}{k} \tag{2}$$

2) Username: On different social network platforms, users usually need to have disparate username to differentiate themselves from others. Different platforms have a different policy about the username. Take Twitter as an example. Twitter has two separate name systems, one called nickname and another called username. The nickname could be the same between different users but the username must be unique. On other platforms, things could be quite different. Regarding as the identity symbol of a user, the username is crucial in user mapping between platforms. To model the username information, we convert the username into a word list.

After lowercase all letters in a username, we could get a list consisted of name words of username. The username of user N on platform A is represented as $U_N^A = (word_1, word_2, , word_j)$.

3) Friend information: According to the social network theory, one usually has overlapped friends on different online social network platforms. If we know a pair of users on two disparate platforms have some overlapped friends, the probability of this pair of users be the same user will be much larger. To modeling the friend information, it is shown as a friend id list. In this paper, we assume that the overlapped users in the dataset are known in the calculating process.

Suppose user N on platform A have n friends $F^A = \left[f_1^A, f_2^A, ..., f_n^A\right]$, and user M on platform B have m friends $F^B = \left[f_1^B, f_2^B, ..., f_m^B\right]$. And they share k overlapped friends $F^o = [f_1, f_2, ..., f_k]$.

4) Location: Most online social network platforms require the user to provide their location information and the location information is usually public. It provides a good opportunity for us to use this information to analysis user similarity. The location information forms can be different among platforms. Some platforms require the location be exact to the city and on other platforms, users could fill the location at their own will. We disassemble the location information into different levels, then directly match the lowest level between the location information pair. Suppose user N on platform A has a location information, we could transform it into a collection like $L_N^A = [City_N, State_N, Nation_N]$. If user N does not have the city information, the first element $City_N$ of L_N^A could be blank, and the rest can be done in the same manner.

C. Similarity computation

Cosine similarity is a measure of similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them. Jaccard index, also known as Jaccard similarity coefficient, is a statistic used for comparing the similarity and diversity of sample sets.

1) User description - Cosine similarity: The vectors we get come from typical words of user description, and they represent the whole meaning of user description. If the two descriptions are similar, the vectors we get will be close in the space. The similarity of user description is calculated using cosine similarity metric.

$$\begin{aligned} Sim(D_N^A, D_M^B) =& \frac{D_N^A \cdot D_M^B}{\left\|D_N^A\right\| \left\|D_M^B\right\|} \\ =& \frac{\sum_{i=1}^{n} D_{Ni}^A D_{Mi}^B}{\sqrt{\sum_{i=1}^{n} {D_{Ni}^A}^2} \sqrt{\sum_{i=1}^{n} {D_{Mi}^B}^2}} \end{aligned} \tag{3}$$

2) Username and Friend - Jaccard index: We could directly use the username list for Jaccard index calculating. Note that the matching of name words is not necessarily identical to each other. If one word is contained in or contains another word, the matching of this pair is valid.

$$J(U_N^A, U_M^B) = \frac{\left|U_N^A \cap U_M^B\right|}{\left|U_N^A \cup U_M^B\right|} \tag{4}$$

The Jaccard index of the pair of users friend information will be:

$$J(F^A, F^B) = \frac{\left|F^A \cap F^B\right|}{\left|F^A \cup F^B\right|} = \frac{\left|F^O\right|}{|F^A| + |F^B| - |F^O|} \tag{5}$$

3) Location - Matching: Directly calculate the similarity using matching. If the lowest non-blank level is the same, $Sim(L_N^A, L_M^B) = 1$. If the lowest non-blank level is not the same, $Sim(L_N^A, L_M^B) = 0$. For example, user A: $L_N^A = [LosAngeles, -, US]$, user B: $L_M^B = [LosAngeles, CA, -]$, the lowest level City is the same (Los Angeles), so the matching is valid. In this case, $Sim(L_N^A, L_M^B) = 1$.

D. Information integration

After calculating the similarity of each element between users, its important to determine the importance of each element, thus we could allot weight for these elements. The weights of each element are expressed as η. The overall similarity calculation method is:

$$S = \eta_d S_d + \eta_u S_u + \eta_l S_l + \eta_f S_f \tag{6}$$

If similarity S is greater than threshold τ, then we consider that the pair of users is the same person.

Through importance analysis, the method could work out for most circumstances. But in the actual scene, the user of the system usually need to tune the weight of each element based on the feedback of the mapping system. This paper provides an interactive interface, allowing users tuning weight of elements and threshold for expecting precision and recall value and adopting more platforms.

IV. Experiment

A. Experiment Settings

The dataset used in the experiment comes from CrossOSN-u public dataset [14]. This dataset contains user information of two online social platforms, Flickr and Twitter, as well as the user connections between the platforms. As the dataset does not provide Twitter username for privacy reason, we crawl the username by Twitter user id using Twitter API ourselves.

There are totally 61733 user data in the CrossOSN-u dataset. To facilitate our framework, 7109 out of them having relatively complete user pair information are selected for the experiment. The fields of the dataset used in the dataset are listed below.

We use the precision, recall, and F1-score for evaluation. The existing user pairs in the dataset are used as the ground truth. In this paper, we use different measuring method of each similarity of elements, and then compare the results with our chosen methods.

B. Experiment Results

The tuning of weights and threshold is a rather difficult and complex process. We first set one weight for independent variable and other weights change together when that one changes. Then gradually change the threshold to get ideal results. The quality of result is measured by F1-score. Fig.2 shows the F1-score curve when changing weights of each

TABLE I
INFORMATION FIELDS

Fields	Note
twitter_id	Twitter user id
flickr_id	Flickr user id
twitter_name	Twitter username
flickr_username	Flickr username
twitter_friendlist	User friendlist on Twitter using twitter_id
flickr_friendlist	User friendlist on Flickr using flickr_id
twitter_location	User location information on Twitter
flickr_location	User location information on Flickr
twitter_description	User description on Twitter
flickr_descritpion	User description on Flickr

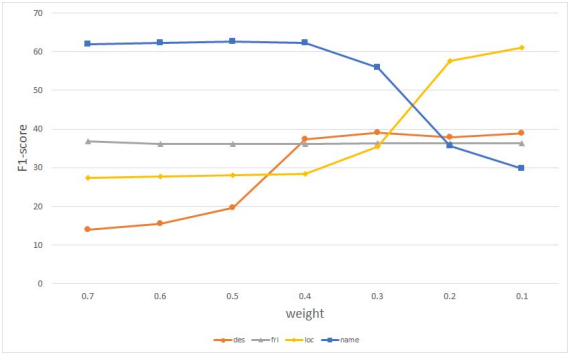

Fig. 2. F1-score changes with different element weights.

element as an independent variable. At the same time, the weights of other elements are changed as dependent variables.

According to the Fig.2, we empirically set the weights and tuning it for the best results. As fig.3 shows, the highest F1 is 62.7 when the threshold τ is 0.36. At this point, the precision is 94.5% and the recall is 46.9%.

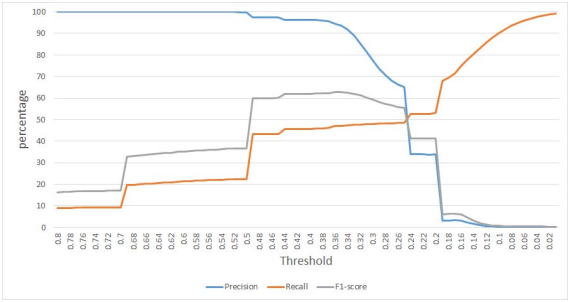

Fig. 3. F1-score, Precision and Recall with optimal element weights.

The result shows that half of the cross-platform user links are identified and recalled, and the precision of the result is relatively high at the same time. When the mapping result is applied in practical work, researchers and regulators usually want a precise result of user mapping rather than recalling a lot of homologous users. So generally speaking, in these circumstances, the precision is more important than the recall.

We compare our methods to other methods to test out the effect of similarity calculation. We change the description similarity calculation method to Jaccard index and friend network similarity calculation method to DeepWalk [15] respectively, and make other methods remain the same. According to the result data, the methods used in this research get higher F1-score than other methods as Fig.4 shows.

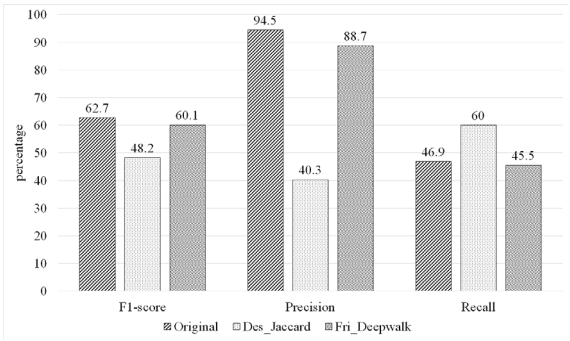

Fig. 4. Result of different similarity calculation method.

C. Integrate user feedback

In this work, we provide a mechanism that allows user to choose the importance of weights for each element, which is useful when the work is conducted practically. For example, the user thinks that the description information deserves more weight than other information, then he could set the weight of description larger. The user could set the expected precision, recall or F1-score simultaneously. Then the system will automatically analyze the information data and reach the expected standard, which is useful in manually doing the related works. The system GUI is shown in Fig.5. According to Fig.5, choosing different weights of elements could lead to different results. The potentially results are listed and with the changing of weights, the overall result could be changed.

D. A case study

Here are three postprocessing user information in the Table.II.

TABLE II
USER INFORMATION

User	A	B	C
Description	Student of life. Designer. Print broker.	Currently a graphic designer An Indefinite student of life.	Content strategy & marketing, publishing and communities. Having fun online.
FriendList	1,2,3,4,5	1,6	1,4,5,7,8
Username	Rob Dupuis	RobDup	rustybrick
Location	New York	Ottawa, Canada	New York, US

After similarity evaluation, the results are shown in the Table.III.

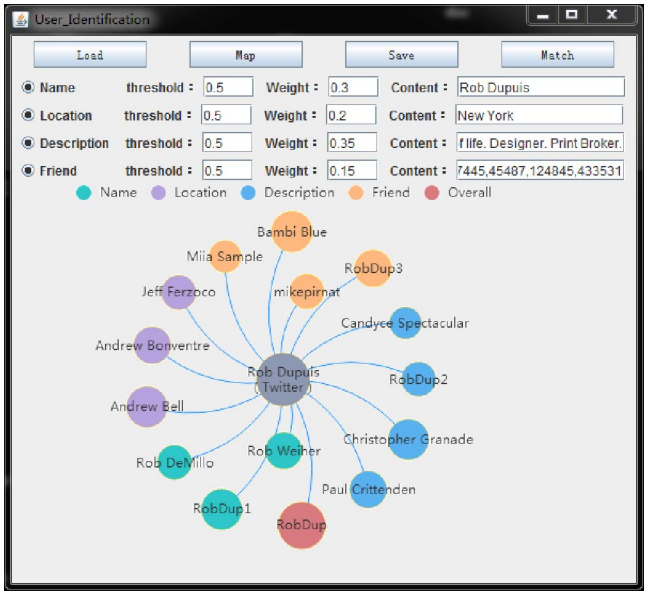

Fig. 5. GUI for user mapping.

TABLE III
SIMILARITY MEASUREMENT

User Pair	A-B	B-C	A-C
S_d	0.857	0.598	0.616
S_f	0.167	0.167	0.429
S_u	1	0	0
S_l	0	0	1
S	0.774	0.196	0.328

According to the results, when we choose threshold $\tau = 0.36$ as we get in the experiment, user A and user B are the most potentially same person. As the two users of different platforms are mapped as the same person, it's more convenient to do the works when it comes to issues such as tracing information sources, tracing rumors sources, skeleton users discovery and doing other security-related works. And researchers could also build recommendation system which could overcome cold start problem as they could get user information from another social platform that this user participates in.

As the parameters can be changed by the people who use this framework in Fig.5, if one chooses high weights for friend and location and low weights for username and description, the result will be totally different with former results. System users could also add more information elements and corresponding weights to perform the user mapping, and get a better model to accomplish the mission of cross social platform user mapping. As we can see, this framework is rather customizable, so users could modify the parameters on their own and make it suitable for majority datasets.

V. CONCLUSION

In this paper, we have introduced the idea of mapping users across social media platforms using user information. It is shown that both text information and structural information are important for user mapping. We proposed a framework using user information which user could modify flexibility of when using in practical work.

Future work includes the completion of missing information using different algorithms. We also expect to find some more useful and effective information to fulfill the user mapping and get better results.

ACKNOWLEDGMENT

This research is supported by the Key Research Program of the Chinese Academy of Sciences under Grant No. ZDRW-XH-2017-3; National Natural Science Foundation of China under Grant No. 71621002,61671450, 71472175, 71402177; National Key Research and Development Program under Grant No. 2016YFC1200702.

REFERENCES

[1] M. Duggan and A. Smith, "Social media update 2013," *Pew Internet and American Life Project*, 2013.
[2] J. Vosecky, D. Hong, and V. Y. Shen, "User identification across multiple social networks," in *Networked Digital Technologies, 2009. NDT'09. First International Conference on*. IEEE, 2009, pp. 360–365.
[3] J. Liu, F. Zhang, X. Song, Y.-I. Song, C.-Y. Lin, and H.-W. Hon, "What's in a name?: an unsupervised approach to link users across communities," in *Proceedings of the sixth ACM international conference on Web search and data mining*. ACM, 2013, pp. 495–504.
[4] L. Liu, W. K. Cheung, X. Li, and L. Liao, "Aligning users across social networks using network embedding," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*, 2016, pp. 1774–1780.
[5] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "word2vec," 2014.
[6] R. Zafarani and H. Liu, "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*. ACM, 2013, pp. 41–49.
[7] T. Iofciu, P. Fankhauser, F. Abel, and K. Bischoff, "Identifying users across social tagging systems." in *ICWSM*, 2011.
[8] G. Kollias, S. Mohammadi, and A. Grama, "Network similarity decomposition (nsd): A fast and scalable approach to network alignment," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 12, pp. 2232–2243, 2012.
[9] X. Kong, J. Zhang, and P. S. Yu, "Inferring anchor links across multiple heterogeneous social networks," in *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. ACM, 2013, pp. 179–188.
[10] T. Man, H. Shen, S. Liu, X. Jin, and X. Cheng, "Predict anchor links across social networks via an embedding approach." IJCAI, 2016.
[11] S. Tan, Z. Guan, D. Cai, X. Qin, J. Bu, and C. Chen, "Mapping users across networks by manifold alignment on hypergraph." in *AAAI*, vol. 14. Citeseer, 2014, pp. 159–165.
[12] Y. Zhang, J. Tang, Z. Yang, J. Pei, and P. S. Yu, "Cosnet: Connecting heterogeneous social networks with local and global consistency," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 1485–1494.
[13] J. Zhang and S. Y. Philip, "Integrated anchor and social link predictions across social networks." in *IJCAI*, 2015, pp. 2125–2132.
[14] M. Yan, J. Sang, T. Mei, and C. Xu, "Friend transfer: cold-start friend recommendation with cross-platform transfer learning of social knowledge," in *Multimedia and Expo (ICME), 2013 IEEE International Conference on*. IEEE, 2013, pp. 1–6.
[15] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: Online learning of social representations," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 701–710.