International Journal of Security and Its Application
Vol.11, No.1 (2017), pp.59-70
http://dx.doi.org/10.14257/ijsia.2017.11.1.06

User Identity Resolution Across Multi-Social Media

Zhibo Wang[1, 2,*], Zhongyuan Li[3], Jiawen Sun[1], Qi Yin[1],
Lu Gao[1] and Xiaohui Cui[1,*]

[1]International School of Software, Wuhan University, Wuhan, China, 430079
[2]Software College, East China Institute of Technology, Nanchang, China, 330
[3]Network Center, Jilin University, Changchun, China, 130023
**{ xcui, rs_wzb } @whu.edu.cn*

Abstract

Nowadays, Internet is fulfilled by huge-scale personal information, and there exists a phenomenon that these kinds of personal information are not distributed in the Internet randomly, but archived in several social media under a certain rule, which is said the different characters of social media. It is rational to infer to the conclusion that we could not learn the every aspect of an individual through only one social medium he/she using, and only if through multi-social media could we learn the more integrated information of individuals. This thesis proposes an approach for user identity resolution based on user's social network and user's behavior patterns. After collecting available user's network and personal information, the approach accomplishes this task that analyze and compare these data, including similarity measure between nodes of networks and string similarity measure, and finally determine whether they are match with each other according to the calculation results. The experimental results show that the method improved the recognition accuracy.

***Keywords**: user identity resolution; user's behavior; social network; multi-social media; information extraction; information fusion*

1. Introduction

Today, Internet users gradually become producers and disseminators of information from the recipient information, with the information generated by Internet growing and Internet community being increasingly opened. This means that Internet users use the network, which is not simply view the display of information, can add new content to the Internet that can contribute new information. Users spread their personal information on the Internet during this period, such information distribute in the network as data. On the contrary, information fusion based on multiple social media is the reverse engineering of the process.

A large number of social media applications, not only brought a lot of data available for analysis, also cause a phenomenon: people may use a variety of social media platforms to publish different information for different purposes or needs. Since some users use a number of different social media platforms to disseminate information, which the user's personal information is divided into blocks scattered among the different social media platforms, rather than being free in all distribution platform.

We can make a reasonable assumption that if we are able to find all the accounts for the same person on different social media platforms, and then analyze the content published on different platforms comprehensively and eventually integrate all these data, we can obtain a more comprehensive summary of personal information.

*Xiaohui Cui is the corresponding author.

Currently, phenomenon that the recommendation system has a single source of information, result in making the recommendation system found that some points of interest of users and recommend information to meet these points by recommending through the data from a single source of information, has some limitations. For example, a point of interest that the user information in this sources do not show has not been found, these interests are also less likely to be met in the recommended News. Thus, it is possible to obtain a more complete analysis of user points of interest and solve the problem of incomplete on recommended content of recommendation system by fusing the user information in multiple social media platforms. Thesis concerns the integration of same user information in different social media platforms, which may have envisioned applications of space in many areas, and can help improve the recommendation system to recommend more accurate target.

2. Related Work

In recent 10 years, there are lots of explorations to match individual identity based on multiple social media platforms, and they are mainly divided into account basic information matching, social relationship matching and the mixed matching.

Data in different social media platforms have similar format, and they are stored in the key/value pair format, which means each key corresponding to a value. It provides convenience for the calculation of similarity. As long as the users’ same item of basic information in two platforms been specified, the calculation of similarity can be carried out. The calculation has stronger operability as for users’ expression information that is relatively messy and has no fixed format [10].

The key to building a social network is the ability of finding people that we know in real life, which, in turn, requires those people to make publicly available some personal information, such as their names, family names, locations and birth dates, just to name a few. However, it is not uncommon that individuals create multiple profiles in several social networks, each containing partially overlapping sets of personal information. Matching those different profiles allows to create a global profile that gives a holistic view of the information of an individual [1].

An increasing amount of many people’s life is spent online. People are using Internet and social media in order to communicate, express their opinions and beliefs, discuss topics of interest to them, *etc*. While much of the information is expressed publicly, there is also more sensitive information available in web forums and other social media services that potentially could be harmful to the author if it became widely known who the physical person behind the user that is posting information is in reality [2].

In the social media platform, users usually pay attention to and be focused on the individuals in real life who share friend or relative relationship with them [11]. These relationships are real, and will show a phenomenon in the social media platform [3]: No matter how different a user’s published content in social media platforms is, the user's social friendship network is relatively stable. A friend existed in user’s one social networking platform, is very likely to appear in the user’s friend list in another social platform. Therefore, according to one account’s friendship, we can describe friendship network of this account and do the similarity calculation with other account’s friendship network.

The proposed method observes user's reactions to the filtered documents and learns from them the profiles for the individual users. Reinforcement learning is used to adapt the most significant terms that best represent user's interests [4].

Recently a solution has been put forward with the combination of the above two methods by some research team, which determines the similarity of two account by the basic information matching and social relationship matching together. This method has better results, because by using two parallel match methods to calculate similarity, it will prevent errors caused by using only one of them [5].

The thesis take the combination measure of user basic information and friend network to implement individual identity matching task based on multiple social media. In this plan, we firstly compute the similarity of social network nodes [12], and then filter out reliable results through user's basic information matching, and finally we will get the conclusion. The innovativeness of this method is reflected in improving the precision of the match helped by user information matching, at the same time exploring network structure to identify the user accounts. We use a single seed user as a starting point, and obtain relevant user's data (limit depth) as the experiment data to complete the experiment procedure. Because the data we obtained cannot be directly used for similarity calculation procedure, we use matrix method to represent friend network. Then, we compute similarity for network nodes and delete unreasonable matching items in the above calculation results by using user basic information as a filter condition. As for the confirmation of results, we will manually lookup to test the result of the experiment.

3. The Hybrid Model based on Combination of Basic Account Information and Social Connection Matching Approach

3.1. Data Acquisition and Assumption

In the respect of access frequency, GitHub is the most relaxed in numerous social website. The default Rate Limit is 60 times per hour and will change to 5000 times per hour after verification. If in this case, you still need to improve Rate Limit, you can negotiate with them through mail, and in that case, you can custom Rate Limit (in the condition of not affecting the normal operation of website).

What is different with GitHub is that some user of Twitter are not physically exist, so these nodes will have effect both on data acquisition and calculation results[13]. We need to specify some filter condition to remove the noise nodes. As the project focuses on ordinary normal person (not including the certification accounts, the promoted accounts, *etc*.), we can specify filter condition as follows: 1, the number of followers or following is more than 6000; 2, the following number is equal to 2001 (due to Twitter limitation); 3, the number of following have an order of magnitude difference between the number of followers at least. In our opinion, you can remove the node as long as you meet above any rule.

3.2. Network Node Similarity Definition

Measure network node similarity:

We used to have an algorithm to measure node similarity in two graphs in order to measure the similarity among users of two social networks of seed users. First, we assume that sum is two directed graphs, the corresponding, and sum is set to have a node collection of two directed graphs. We define a matrix S of order, each element of the matrix S represents the degree of similarity between node j in graph A and node i in graph B, called similarity values.

Kleinberg proposed a method to distinguish between a focus for a keyword query results, which is good Hubs, which is good Authorities. For example, for "university" query, Oxford, Harvard or other University's home page, known good Authorities, those who point to the home page called good Hubs. Good Hubs point to good Authorities, at the same time good Authorities are pointed by good Hubs. From this implicit relationship, we can make each node in a given figure using an iterative method assignment (Hub Authority value and value).

We have assumed that the point set V and edge set E, and defined sum equals Hub and Authority values of node j. First, we give each node two non-negative, then update both the properties of these nodes by the following procedure: Hub value of node j is equal to the sum of Authority value of nodes which node j refers to, the same, Authority value of node j is equal to the sum of Hub value of nodes which node j refers to. Indicate the following:

$$\begin{cases} h_j \leftarrow \sum i : (j,i) \in E^{a_i}, \\ a_j \leftarrow \sum i : (i,j) \in E^{h_i}, \end{cases} \tag{1}$$

We use graph B represents graph G's adjacency matrix, h represents vector consisting of Hub values and a represents vector consisting of Authority values, then the updating process can use the following simple form, it said:

$$\begin{bmatrix} h \\ a \end{bmatrix}_{k+1} = \begin{bmatrix} 0 & B \\ B^T & 0 \end{bmatrix} \begin{bmatrix} h \\ a \end{bmatrix}_k, k = 0,1,..., \tag{2}$$

Further simplification：

$$x_{k+1} = Mx_k, k = 0,1,..., \tag{3}$$

among them:

$$x_k = \begin{bmatrix} h \\ a \end{bmatrix}_k, M = \begin{bmatrix} 0 & B \\ B^T & 0 \end{bmatrix}, \tag{4}$$

Because we are only interested in the relative coefficient and absolute coefficient does not make sense, we need to normalize the results:

$$z_0 = x_0 > 0, z_{k+1} = \frac{Mz_k}{\|M\ z_k\|_2}, k = 0,1,..., \tag{5}$$

We used z_k as the final result, but there are two issues which need to be solved, first, initial value problem, and second, convergence problems. Concerning the initial problem, we decided to use a matrix which all the elements equal to one as the initial value. With regard to convergence problems, above-mentioned iterative process is not always convergent, but it's even or odd sequence always convergent, we choose one.

Now we have to promote this process of generalization: Authority value of node j in graph G can be understood as similarity values in G's node j and the Authority node. Similarly, Hub value of node j in graph G can be understood as similarity values in G's node j and the Hub node.

It can be simplified as follows:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}_{k+1} = \begin{bmatrix} 0 & B & 0 \\ B^T & 0 & B \\ 0 & B^T & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}_k ,\text{k = 0,1,...,} \tag{6}$$

Further simplification：

$$x_{k+1} = Mx_k, k = 0,1,2,... \tag{7}$$

Now we will further generalize the above reasoning, assuming that there are two pictures:

Graph A serves as the example among 1 → 2 → 3 role and we define the result as a matrix S of order $n_B \times n_A$, each element x_{ij} of the matrix S is expressed as the value of attribute j of node i in graph B, above-mentioned iterative update process can be summarized by the following formula:

$$X_{K+1} = BX_K A^T + B^T X_k A, k = 0,1,..., \tag{8}$$

Among them, A, B is the adjacency matrix of graph A and graph B respectively. We use the limit value of the above-described iterative process as a similarity matrix definition.

Similarity matrix calculation algorithm is described as follows:

1, initialize $z_0 = 1$;

2, make this formula iterate the even times until limit is reached;

3, the limit of the iterative process described above is similarity matrix to the graph A and graph B.

4. Experiment

We use a combination of basic account information and social connection matching approach to match the identity of the individual, where the first is social connection matching, and then is individual identity matching to filter.

4.1. Basic Account Information Match

Basic account information includes all publicly available data on the user's social networking platform, including a user's profile information, such as user name, homepage, mail, companies and *etc*. and user tweets text messages. Then the user text information will be preprocessed, entity recognized, extracted with related words and information matched.

4.1.1. Text Preprocess: 1) Cutting text words: can cut user social text word and the processing result is the processing word set.

2) words standardization: can standardize the result of cutting text words in a different voice and form for the same word, and the processing result is the basic set of words or word radical in order to reduce the workload of the subsequent steps.

3) Text POS tagging (POS): can tag the result of cutting text words and words standardization with text annotation, POS tagging word processing result after collection.

4.1.2. Text Entity Recognition: Text entity recognition function can extract information from social text, which can extract regular entity information from short social text without rules, which means that entity recognize for the specified user’s all social text based on the specified unique identifier in social network, and the processing result is the collection of entities and entity types [14].

4.1.3. Entity’s Related Words Extraction: Entity’s related words extraction function can extract information from social text, especially extract regular entity information from short social text without rules. Here we use two methods for extracting entity’s related words, they are sentence method and the chi-square method [8].

1) Sentence Method:

Procedure: 1. Classify the sentence pattern according to the various parts of speech in a sentence; 2. Get word entity’s related words extraction rule of this sentence according to matching with the trained sentence library; 3.Read the sentence entity word POS (position) and search for the related words of what part of speech (which some locations) at the POS (position)’s entity in the relevant word extraction rules; 4. Take out words that appear before a words as the relevant word according to the frequency of related word at each position, n and a is average of related word.

2) Chi-square method:

Procedure: 1.Obtain an entity word E in a text, and obtain the other words w in the text (not including the stop words); 2.Count the following four statistical data throughout the anticipated library: text number n11 when E appears and w appears, text number n10 when E appears and w does not appear, text number n01 when E does not appear and w appears, text number n00 when E does not appear and w does not appear; 3.Calculate the probability of each event which are e11, e10, e01, e00, formula are shown below;

4.Calculate the chi-square statistic X^2 of E and w, formula are shown below; 5.If X^2 equal or greater than the chi-square critical value 10.83 when confidence level is 0.999, then w and E are not independent, that w is related words to E. The confidence level value 0.999 on behalf of E and w has 99.9% probability related, and the possibility that they are not related is almost equivalent to the probability of low probability events; 6.Calculate the result of E with chi-square statistic for each word of the text and determine whether they are related according to step 1-4.

e11 formula (e10, e01, e00 formula so on):

$$e_{11} = n * \frac{n_{11} + n_{10}}{n} * \frac{n_{11} + n_{01}}{n}, n = n_{11} + n_{10} + n_{01} + n_{00} \tag{9}$$

X^2 formula:

$$x^2 = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} \frac{(n_{e_t} n_{e_c} - n_{e_t} n_{e_c})^2}{n_{e_t} n_{e_c}} \tag{10}$$

4.1.4. Basic Account Information Match: For the text entity in twitter and github user's basic information, which are listed separately collection of entity types, such as a collection of people, organizations, places, and so on. Calculate the similar values which is similar values weighted average value of each entity set in these entities on two social networking platforms by SOUNDEX algorithm. That is:

$$m_{gt} = \frac{\sum_{i=0}^{n} e_i x_i}{n} \tag{11}$$

For github user g and twitter users t, M is the two accounts Information match value, n is the number of the entity type, e is weight for each entity class, x corresponds to user g and t 's similarity at this entity type.

4.2. Social Relationship Matching

In the social media platform, users usually pay attention to and be focused on the individuals in real life, who share a friend or relative relationship with the user [15] [6]. These relationships are real, and will show a phenomenon in the social media platform: No matter how different a user's published content in social media platforms is, the user's social friendship network is relatively stable [9][16]. A friend existed in user's one social networking platform, is very likely to appear in the user's friend list in another social platform. Therefore, according to one account's friendship, we can describe friendship network of this account [17] and do the similarity calculation with other account's friendship network.

Hypothesis: If there exist high network structure similarity between two nodes, we have full confidence that the two nodes is the same entity.

4.2.1. Measure Similarity of Network Nodes: For the similarity calculation of nodes in the network, we proposed to use the calculation method which is provided by Vincent D. Blondel[7] and others. We take the way to calculate the similarity between two network nodes which use adjacency matrix representation. The calculation process is an iterative process, and the results will be represented by using matrix. The initial result default as one, and then run iterative times until the result matrix convergence reach limits. We set a threshold value, and we regard the nodes pair whose similar degree is over the threshold as the same person. Then we can delete obviously unreasonable matching results through the user basic information matching.

4.2.2. Test based on Simple Seed Node: From the a simple seed node (identified by the same individual, identification standard is as below) and using the user account information on Twitter and GitHub, we respectively obtain the user's relationship network of two layers in these two social platforms and the corresponding similarity between two network nodes, and draw out diagram of the relationship between two network, as shown in Figure 1 and Figure 2, which is respectively the relationship network of Twitter and GitHub node. We can calculate a sorting list of nodes that have the maximum similarity, and with the manual inspection we can get recognition accuracy.

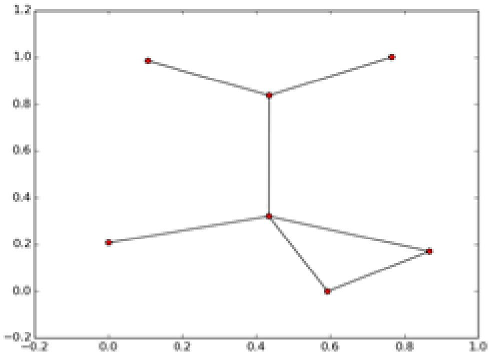

Figure 1. Simple User's Twitter Network

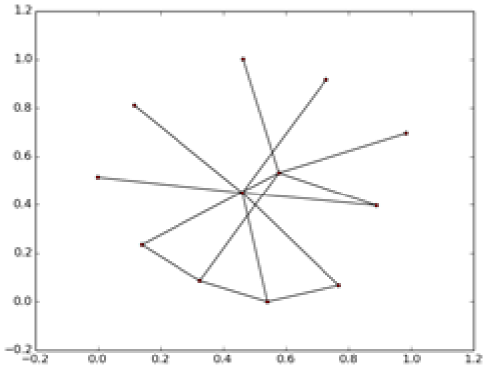

Figure 2. Simple User's Github Network

Artificial inspection standards: for the two nodes of different social networking platforms, we decide whether it is the same individual comprehensively according to the following conditions: 1. The head portrait similarity; 2. User name similarity; 3. The home page platform correlation (that means the home page links to other social platforms). 4. Content similarity. After inspection, identification accuracy can up to 10%.

4.2.3. Test based on Ordinary Seed Node: The network relationship of ordinary users basic follows power-law distribution, and many public figures and those users who are not exist in the physical world will have great influence on the experimental results. So we need to specify the corresponding filters to remove these noise nodes.

Filter condition is as follows: 1, the number of followers or following is more than 6000; 2, the following number is equal to 2001 (due to Twitter limitation); 3, the number of following have an order of magnitude difference between the number of followers at least. In our opinion, you can remove the node as long as you meet above any rule. Network is as shown in Figure 3, Figure 4:

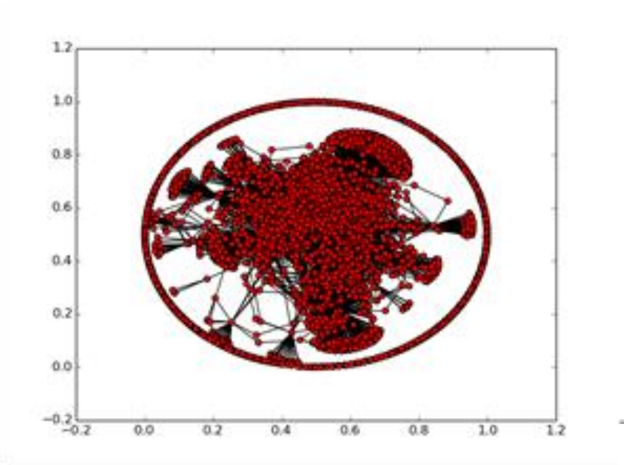

Figure 3. Normal User’s Twitter Network

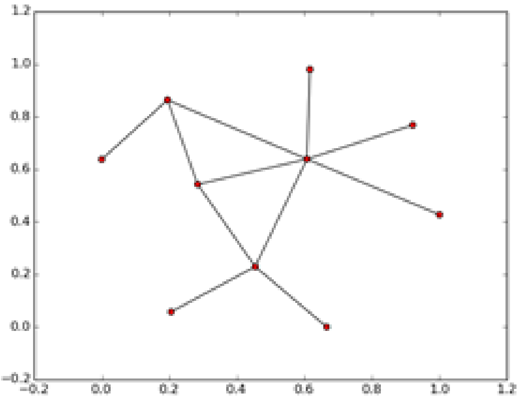

Figure 4. Normal User’s Github Network

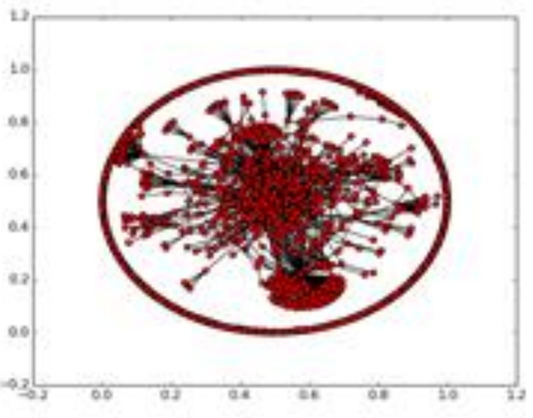

Figure 5. The First Iteration Result of Ordinary Seed Node Twitter Social Relationship

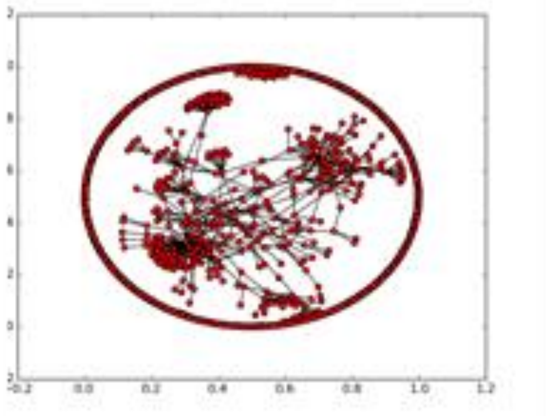

Figure 6. The Second Iteration Result of Ordinary Seed Node Twitter Social Relationship

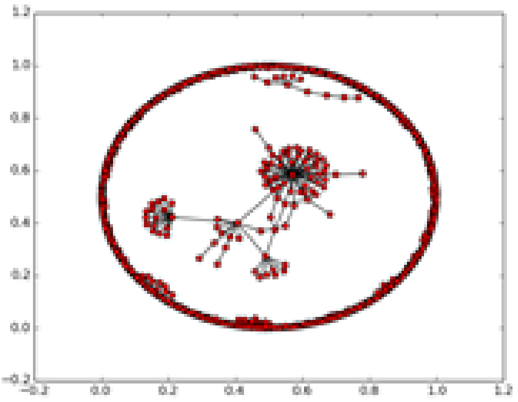

Figure 7. The Third Iteration Result of Ordinary Seed Node Twitter Social Relationship

We adopt the method of quadratic iterative filtering for further noise reduction and processing for the raw data on Twitter. The basic process of secondary screening: 1. Using the similarity algorithm showed in 5.2.1 figure to calculate the similarity of each node and get a list for similarity; 2. Set the threshold, retain the node whose similarity is higher than the threshold value to get new relationship network; 3. Repeat steps 1 and 2. Iteration results in Figure 5, Figure 6 and Figure 7.For ordinary seed node, the similarity value we calculate is low, and recognition result is weak, so we define node related rate to evaluate the reliability of his algorithm applied in the social network. The definition of node related rate:

For two nodes with comprehensive consideration of the following conditions, we have abundant reasons to think they are related: Can be judged as the same individual through artificial inspection standard; For node on Twitter platform, can artificial test to identify GitHub node from the fans and followers list, and vice versa. Node related rate equals to the result of the number of related nodes divided by the number of matching nodes*100%.

According to the definition, we give the following related rate statistical table after the number of iterations.

Table 1. Related Rate Statistics

	Twitter nodes	Relative nodes	The node rate
Initial input	27120	653	2.408%
The first iteration	14232	497	3.492%
The second iteration	6754	265	3.923%
The third iteration	1523	132	8.643%

We can see that the twitter’s initial node order of magnitude difference is very big comparing to github, so the node related to the rate of inspection is low. The iteration results can be got through the threshold filter out irrelevant nodes, and it can be seen after the iteration screening node related rising rates in small increments. After three iterations screening, we get the node related rate and is only 8.643%.

5. Conclusion

In identity matching based on multiple social media platforms, we combine the basic account information matching and social relations matching method to match user information. In match of the user basic information, we use natural language processing to extract the user's entity information. According to the similarity between the entities of the user to determine similarity between two user nodes. In match of social relations, we apply graph similarity algorithm created by Vincent D.Blondel, Anahi Gajardo, Maureen Heymans, Pierre Senellart, and Paul Van Dooren to Twitter and Github social network, respectively, under the simple nodes and node mode, and then compare the similarity of the two networks. Due to the two social networks have unequal size, in order to obtain more reliable match results, our iterative filter method deletes noise nodes with a low relation similarity. For a simple node, it fits a higher degree of user similarity. There is 10% probability of recognition to the same individual in different platforms. For normal nodes, it fits a lower degree of user similarity, the corresponding probability is only 8.643%.

Acknowledgements

This research is supported in part by National Nature Science Foundation of China No. 61440054, Fundamental Research Funds for the Central Universities of China No. 216274213, and Nature Science Foundation of Hubei, China No. 2014CFA048.

Outstanding Academic Talents Startup Funds of Wuhan University, No. 216-410100003. National Natural Science Foundation of China (No. 61462004). Natural Science Foundation of Jiangxi Province (No. 20151BAB207042). Youth Funds of Science and Technology in Jiangxi Province Department of Education (No. GJJ150572).

References

[1] Bennacer, N., Jipmo, C. N., Penta, A., & Quercini, G. (2014, June). Matching User Profiles Across Social Networks. In International Conference on Advanced Information Systems Engineering (pp. 424-438). Springer International Publishing.

[2] Johansson, F., Kaati, L., & Shrestha, A. (2015). Timeprints for identifying social media users with multiple aliases. Security Informatics, 4(1), 1.

[3] Krause, J., Croft, D. P., & James, R. (2007). Social network theory in the behavioural sciences: potential applications. Behavioral Ecology and Sociobiology, 62(1), 15-27.

[4] Levchuk, G. M., Lea, D., & Pattipati, K. R. (2008, April). Recognition of coordinated adversarial behaviors from multi-source information. In SPIE Defense and Security Symposium (pp. 694305-694305). International Society for Optics and Photonics.

[5] Ma, Y., Zeng, Y., Ren, X., & Zhong, N. (2011, September). User interests modeling based on multi-source personal information fusion and semantic reasoning. In International Conference on Active Media Technology (pp. 195-205). Springer Berlin Heidelberg.

[6] Qian, H. (2008). Social relationships in blog webrings. ProQuest.

[7] Blondel, V. D., Gajardo, A., Heymans, M., Senellart, P., & Van Dooren, P. (2004). A measure of similarity between graph vertices: Applications to synonym extraction and web searching. SIAM review, 46(4), 647-666.

[8] Wagner, W. (2010). Steven bird, ewan klein and edward loper: Natural language processing with python, analyzing text with the natural language toolkit. Language Resources and Evaluation, 44(4), 421-424.

[9] Veldman, I. (2009). Matching profiles from social network sites: Similarity calculations with social network support.

[10] Raad, E., Chbeir, R., & Dipanda, A. (2010, September). User profile matching in social networks. In Network-Based Information Systems (NBiS), 2010 13th International Conference on (pp. 297-304). IEEE.

[11] Johansson, F., Kaati, L., & Shrestha, A. (2013, August). Detecting multiple aliases in social media. In Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (pp. 1004-1011). ACM.

[12] Liu, S., Wang, S., Zhu, F., Zhang, J., & Krishnan, R. (2014, June). Hydra: Large-scale social identity linkage via heterogeneous behavior modeling. In Proceedings of the 2014 ACM SIGMOD international conference on Management of data (pp. 51-62). ACM.

[13] Zhao, W. X., Jiang, J., He, J., Song, Y., Achananuparp, P., Lim, E. P., & Li, X. (2011, June). Topical keyphrase extraction from twitter. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1 (pp. 379-388). Association for Computational Linguistics.

[14] Norvig, P. (1987, July). Inference In Text Understanding. In AAAI (pp. 561-565).

[15] Vosecky, J., Hong, D., & Shen, V. Y. (2009, July). User identification across multiple social networks. In 2009 First International Conference on Networked Digital Technologies (pp. 360-365). IEEE.

[16] Domingos, P. (2004). Multi-relational record linkage. In In Proceedings of the KDD-2004 Workshop on Multi-Relational Data Mining.

[17] Soltani, R., & Abhari, A. (2013, July). Identity matching in social media platforms. In Performance Evaluation of Computer and Telecommunication Systems (SPECTS), 2013 International Symposium on (pp. 64-70). IEEE.

Authors

Zhibo Wang Ph.d candidate, International School of Software, Wuhan University. Major in Big data analysis, data mining, *etc*.

Xiaohui Cui Professor, doctoral supervisor, International School of Software, Wuhan University. Major in Big data analysis, High performance computing, Cloud computing, Network security, GPS computing, *etc*.