International Journal of Research in Advent Technology, Special Issue, March 2019
E-ISSN: 2321-9637
Available online at www.ijrat.org

User Identification across Social Media based on User Generated Cross- Link Posts

Waseem Ahmad[1], Rashid Ali[2]
Department of Computer Engineering[1]
AMU, Aligarh[1]
India,202002

Abstract-An online user shares his personal identities like name, username, email address and contact details, etc. during online social profile creation. As each user shares a certain set of attributes on a particular social media service, the amalgamation these publicly available attributes may be utilized to build a complete profile of the user, which can further be exploited in product recommendation, online social marketing, crime influence analysis, etc. To find the user attributes similarity in intra and inter social network environment, we propose a framework which exploits the publicly available cross-linked post present on the social networking site Twitter and extracts the user's profile and network information to determine the correlation among these features. Experimental results show that around 27% of Twitter users have chosen the same screenname and fullname, while the accuracy of exact username matching across the networks is about 92%.

Keywords : The social network, Personal information, Account matching, Online conversation.

1. INTRODUCTION

An online user creates accounts on different social media to interact with his friends and relatives. Each social media is intended for its popular services, for example; Instagram for real-time visual information like photos and images, Twitter for recent news and Facebook contain personal data like birthday, posts, etc. Such social media generates revenue by mining the user-generated content like user interest, preference, the area of expertise, etc. For instance, we are getting ads about favorite items, movies, travel related, tours and travel-specific advertisements on our email or even on the personal mobile too as recommendations. One of the significant challenges is how such information is recommended to a particular user without his awareness. In the background, many such activities are performed by the business organizations to understand the behavior of the user communities by analyzing users shared information on social media services.

User identification across the social networks may help in understanding the behavior and interest of the people, which may further be exploited by the business and advertising agencies to recommend a particular item to a user on his/her email id and personal mobile phone. Such recommended information can be used to improve the business by sharing the personalized messages within budget and economical as compared to the other advertising agencies like TV channel and Newspapers etc.

Many online users often reveal their social account information during an online conversation. Such information is classified into username based information leakage and URLs based information leakage. Public social media like Twitter provide information to its researchers for getting useful information by analyzing users publicly shared contents either in text or visual form. In this work, we retrieve user publicly shared information on Twitter by using Twitter's API thereafter, we extracted screen names and name to understand the similarity among the usernames and real-name. In the second phase, we matched the user-identity simply based on screen-name and shared a cross-linked post on Twitter.

The main contribution of our works is as follows:

- Firstly, we automatically extracted users' shared information from Twitter via Twitter API by developing a query which gives cross network information like "follow me on Instagram".
- Then we filtered the screen-name and real-name from the unstructured information.
- After that, we matched the screen-name and real-name of the users present on the Twitter extracted by using only the following relationship available on Twitter.
- Finally, we match the screen names of the corresponding to the social networks, a keyword like Instagram, Facebook, and Pinterest. For matching, we only used the Levenshtein distance measure. Such a result may help in finding the users' relationship across social networks by utilizing their network relationships.

2. RELATED WORKS

Online social account matching is a well-investigated problem, in literature it is known by different names; social identity matching [1], cross-platform social identity resolution [2] [3], social account mapping [4], user identification across multiple social networks [5]. Social network-based information can be classified into three main categories; profile-based personal information, content-based personal information and the network-based personal information [6] [7]. During the initial stage of social network research, researchers and practitioners prefer profile-based information to find identical users across different social networks. Profile-based information includes username, real name, gender, location, birth-date, contact numbers and addresses, such information often filled by users during the online profile creation process. In paper [5], the authors used profile attributes, name, username, and location, to find the identical users across Facebook and StudiVZ[1], while Goga et al. [7,11], utilized profile attributes real name, username, location, and profile picture to find identical users across Twitter, Facebook and Google+[2]. Malhotra et al. [8], applied data mining techniques on two popular social networking sites Twitter and LinkedIn. They suggested that the username and name are the most discriminating features of the profile attributes or user identification. Peled et al. [9], focused on profile-based features like a real name, username, location, gender they applied machine learning techniques to find the same user on Facebook and Xing with better accuracy.

In the middle stage of the research, people utilized the content-based approach to similar account across social media. The content-based approach includes users writing style, posted content, and location-based features. To identify a user, Zheng et al. [10] have given a framework which describes the use of the writing style of the user as a useful attribute to find a user as identical. Later, Goga et al. [11], used Geo-location based features present in the in the tweets. Almishasri and Tsudik [12] also described a method to identify a user using the writing style of the users. More recently, researchers have focused their interest towards network-based information. Narayan and Shmitikov [13] exploited a network-based approach to de-anonymize users across different social networks. Bartunov et al. [14] utilized the joint link attributes using a conditional random field to find identical users across Facebook and Twitter.

Further, Korula and Lattanzi [15] proposed an efficient reconciliation algorithm for user identification across social networks. More recently Zhou et al., proposed a dynamic framework to identify users across social media networks using friend relationship-based approach. Some of the researchers exploited hybrid techniques to identify users, in this context, Iofciu et al. [16] combined profile attributes; name and username with content attribute user tag to find identical users across the social tagging system. Jain and kumaraguru [3] make use of user profile, content and network attribute to resolve user's identity across social media services.

3. PROPOSED FRAMEWORK

In this paper, we present structural design for user's identity across social networks using cross-linked posts. Let P be the posts shared by users on social network T (Twitter). Then the number of posts obtained through a query Q be m and is represented by N_i. It can be represented by (1).

$$N_i=Q(T(P)) \quad (1)$$

where, $1 \le i \le m$

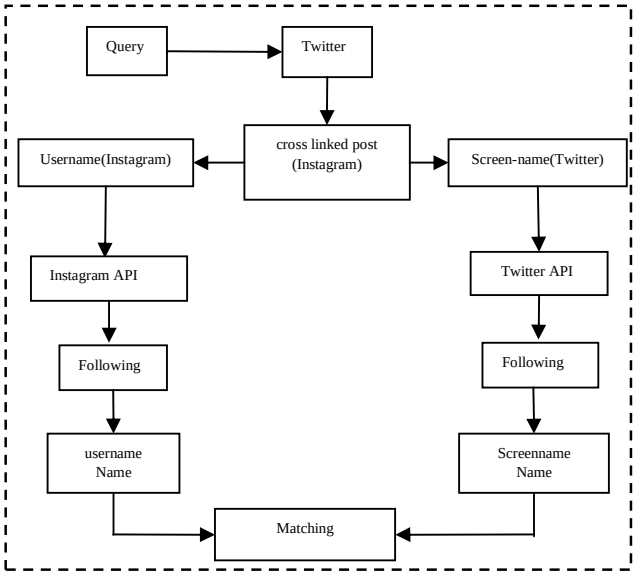

Figure 1: Information Extraction using cross link posts.

(a) Screen-name and name matching

Twitter is a text-rich microblogging social network, which contains both screen-name and real name to represent a user along with profile pics. In this work, first, we understand the user's identity, the similarity in an intra-network manner, after that, we check user identity similarity across the social networks. To match the user's identity, we use three similarity measures, namely Levenshtein, distance measures[17].

(b) Screen-name matching across the social networks

In this technique, we only utilized the username or screen-name to match the user's identity across the sites. Such a minimum number of attributes may help find large numbers of identical users on two syntactically similar social networks like Instagram and Twitter. To match the

user's identity across the social networks, we only users Levenshtein distance [17].

4. DATA COLLECTION AND EXPERIMENTAL SETUP

To prepare a real-world dataset, we build a set of queries like "follow me on Instagram", "Find me on Instagram" etc. to retrieve the relevant, personalized information from Twitter. Thereafter, we retrieve a set of results and randomly selected only three posts to find the users as seed users. Thereafter, we exploit seed attributes to extract users' network information like an outgoing relationship (Following relationship) matching these attributes we find that a single seed user can contribute around 55 now seed users for further iteration. By using initially obtained three users, we extracted only outgoing relationship. The extracted data obtained is depicted in table 1.

Table1. Dataset: Number of users on the social network

S. No.	Social Network	Number of Following connections
1	Twitter	188215
2	Instagram	55245

From table 1, it is found that for an ordinary social accounts number users on Twitter is more than Instagram. Because Twitter's friendship network is unidirectional while for Instagram it is bidirectional.

5. RESULT AND DISCUSSION

(a) Intra-network Identity Matching

To find the users identity similarity in the intra-network environment, we matched the user's fullname and screen-name by applying Levenshtein distance. We performed three successive experiments by selecting 10000 pairs. From the observation of table 2, it is found that around 27% of users have chosen the same fullname and username on Twitter. It is obtained that most of the people on Twitter do not share the real name. But 27 percent user put the same username/screen name with a real name; we assumed them as authentic profile to match the users across the social networks. Further, it is observed that those people who have one word username or more similar than the larger username size.

Table 2. Name and screen-name matching on Twitter

S. No.	Twitter	Number of attributes	Matched (Levenshtein Distance)
1	Name	10000	2412
	Screen-name	10000	
2	Name	10000	2716
	Screen-name	10000	
3	Name	10000	3016
	Screen-name	10000	

(b) Internetwork Identity Matching

We randomly selected forty seed users and accessed their following relationship and matched the username across the social networks by using only Levenshtein distance. For seed profiles contain 33227 users on Twitter while 27445 users on Instagram. Matching these users' identity (username) using Levenshtein distance we found that 1456 users have exactly the same username while 443 users have a similar username with one character in difference i.e distance is one. The results of the proposed framework presented in equation Table. 3.

Table 3. Username based account matching results

S. No.	Social Network	Number of Accounts	No. of account with Levenshtein score	
			Zero	One
1.	Twitter	33227	1556	443
2.	Instagram	27445		

From the observation of table 3 it is obtained that 1556 users have the same names across the site. Further, we evaluated the results in the real environment it is obtained that 91.35 % users have identical account across two popular social media services like Twitter and Instagram. When we evaluated distinct username, then find the accuracy around 45 % with a distance score one.

6. CONCLUSION

In this research article, we proposed a framework to match users account relevant information across social networks. This framework utilizes the cross-link posts present on Twitter and finds the seed users. Then, these seed users to obtain network relationships. Further, we matched users' name and username present in the intra-network and

internetwork environment. the user's identity on the same network and across the networks. We obtained that our proposed method shows the 92% accuracy with zero scores and 45% similarity with one score.

REFERENCES

[1] J. Li, G. A. Wang, and H. Chen, Identity matching using personal and social identity features. Information Systems Frontiers, vol. 13, no. 1, pages 101-113, 2011.

[2] K. Shu, S. Wang, J. Tang, R. Zafarani, and H. Liu, User Identity Linkage across Online Social Networks: A Review. ACM SIGKDD Explorations Newsletter vol. 18, no. 2: pages 5-17, 2017.

[3] P. Jain, P. Kumaraguru and A. Joshi, @ i seek'fb.Me': Identifying users across multiple online social networks. In Proceedings of the 22nd international conference on World Wide Web ,pages 1259-1268, 2013.

[4] 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, pages 159-165. 2014.

[5] J. Vosecky, D. Hong, and V. Y. Shen, User identification across multiple social networks. In First International Conference on Networked Digital Technologies, pages. 360-365. 2009.

[6] X. Zhou, X. Liang, H. Zhang, and Y. Ma, "Cross-platform identification of anonymous identical users in multiple social media networks," IEEE transactions on knowledge and data engineering, vol. 28, no. 2, pages.411-424, 2016.

[7] O. Goga, , D. Perito, H. Lei, R. Teixeira, and R. Sommer, Large-scale correlation of accounts across social networks. University of California at Berkeley, Berkeley, California, Tech. Rep. TR-13-002, 2013

[8] A. Malhotra, L. Totti, W. J. Meira, P. Kumaraguru, and V. Almeida, Studying user footprints in different online social networks. In International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pages 1065-1070, 2012 .

[9] O. Peled, M. Fire, L. Rokach, and Y.Elovici,. Entity matching in online social networks. In International Conference on Social Computing (SocialCom), pages 339-344, 2013

[10] R. Zheng, H. C. Li, and Z Huang., A framework for authorship identification of online messages: writing-style features and classification techniques. J. of the American Society for Information Science and Technology, vol. 57, no. 3, pages. 378-393, 2006.

[11] O. Goga, H. Lei, S.H.K. Parthasarathi, G. Friedland, R. Sommer, and R.Teixeira, Exploiting innocuous activity for correlating users across sites.In Proceedings of the 22nd international conference on World Wide Web, pages 447-458, 2013

[12] M. Almishari and G. Tsudik, Exploring linkability of user reviews. Computer Security–ESORICS pages 307- 324, 2012.

[13] A. Narayanan and V. Shmatikov, De-anonymizing social networks. In Proceedings Of the 30th IEEE Symposium on Security and Privacy, pages 173-187, 2009

[14] S. Bartunov, A. Korshunov , S. Park, W, Ryu., and H. Lee, Joint link-attribute user identity resolution in online social networks. The 6th SNA-KDD Workshop, 2012

[15] N. Korula, and S. Lattanzi, An efficient reconciliation algorithm for social networks.In proceedings of the VLDB Endowment, vol. 7, no. 5, pages377-388, 2014.

[16] T. Iofciu, P. Fankhauser, F. Abel, and K. Bischoff, Identifying users across social tagging systems. In ICWSM, 2011.

[17] W. Cohen, P. Ravikumar, and S.Fienberg, A comparison of string metrics for matching names and records. In Kdd workshop on data cleaning and object consolidation Vol. 3, pages. 73-78, 2003