National Conference on Recent Advances in Computer Science and IT (NCRACIT)
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 4 | Issue 1 | ISSN : 2456-3307

A Framework for User Identity Resolutions across Social Networks

Suhail Iqbal Bhat[1], Tasleem Arif[2], Majid Bashir Malik[3]

[12]PG Department of IT, BGSB University, Rajouri, Jammu and Kashmir,India
[3]Department of Computer Sciences, BGSB University, Rajouri, Jammu and Kashmir, India
iqbalsuhail23@gmail.com[1],tasleem.ap@gmail.com[2],majid.malik@rediffmail.com[3]

ABSTRACT

Today, over 2.62 billion people are active social media users accounting to one-third of the world's population. There exist hundreds of online social networking sites offering different services and functionality to their fellow user, however few among them are the most popular like Facebook, Twitter, Instagram and LinkedIn. Social networks are designed to address specific social needs, offering a distinct set of services and functionality. In order to enjoy the diverse range of services and to cover different facets of life, a user often registers on multiple social networks resulting in dissimilar identities of same users. The process of finding and linking those similar but disconnected identities of an individual scattered across social networks is termed as Identity Resolution or User identity Linkage. It has a significant impact on various problem domains such as recommendations, target marketing, user profiling, impersonator detection, etc. In this paper we propose a framework to the identity resolution problem. We also discuss various challenges of linking user's identities across online social networks.

Keywords : Online Social Network(OSN); Identity Resolution; Profile; Machile Learning; Facebook; Twitter

I. INTRODUCTION

As on January 2018 more than 3.87 billion people have access to Internet and 2.62 billion of them are active Social Media users accounting to one-third of the world's population. Facebook, the most popular and widely used has over 2,167 million active users followed by 800 million on Instagram, 330 million on Twitter, 106 million on LinkedIn [27] . In recent years, online social networks have emerged as a new object of study. We are witnessing the increasing participation of people in online activities like content posting and having different kinds of interactions and relationships along with the rise of online social networks and the extensive tendency of people toward them. This has resulted in generation of and making available huge volume of valuable data that was never available before, which in turn introduced some new, attractive, varied, and useful research areas to researcher community [26]. Analyzing large scale social networks is used extensively in a wide range of applications and disciplines. Some of the applications include data aggregation and mining network modeling, user attribute and behavior analysis, community detection and analysis, rumor identification, development of recommender systems, link prediction and identity resolution.

Social networks have transformed the Internet ecosystem leading to a more collaborative environment. The services provided by social networks have changed how we approach many facets of life, it has changed the way we used to communicate with others, entertain, educate and actually live. There are hundreds of online social networks available today designed to address specific social needs and offering a distinct set of services and functionality [22]. Some of the social networks have a general audience like *Facebook* and *Whatsapp*

whereas others have a specific audience like *Twitter* which is popular among users who want to share their opinions and quick access to news, *Instagram* among Youngsters to share their leisure activities, *Pinterest* among Artists to showcase their art and *LinkedIn* among the Professional to connect to employers. In order to enjoy the diverse range of services and to cover different facets of life a user often registers on multiple online social networks resulting in disparate unlinked identities scattered across the social networks [10]. A study conducted by the Pew Research Center shows overlap of users with one another among five popular social networks: *Facebook, Twitter, Instagram, Pinterest* and *LinkedIn* [20]. It can be seen from the Table-1 that high correlation among the users is found in *Facebook, Twitter* and *Instagram*. 91% users resister themselves both on *Twitter* and *Facebook*; 58% users on *Twitter* and *Instagram.* The pace with which the social media users are growing and social networks are enhancing their services to attract more users, this overlap is expected to increase more.

Table1: Social Media Matrix

Social Network	*Use Twitter*	*Use Instagram*	*Pinterest*	*LinkedIn*	*Use Facebook*
% of Twitter Users who	-----	58 %	42%	47%	91%
% of Instagram Users who	52%	-----	47%	38%	94%
% of Pinterest Users who	34%	43 %	-----	40%	88%
% of LinkedIn Users who	39%	35 %	40%	-----	86%
% of Facebook Users who	29%	34 %	34%	33%	----

There is a rising interest in identifying multiple accounts that correspond to a single individual across the social network and within a single social network. First, drawing a link between the unlinked identities scattered across social networks can be helpful in many scenarios such as user profile integration, business intelligence, cyber forensic and national security [22]. In addition to this matching the accounts that correspond to same individual can be helpful to the e-commerce sites by providing better personalized recommendation and target marketing to their customers. Second, social networks are interests in finding all accounts corresponding to single individual inside a single social network. These accounts are used by malicious minded people as a platform to execute ill activities like cyber bullying, impersonation, spamming, identity deception, etc. For both cases, we need matching techniques to find the accounts of a single individual.

An identity of a user on an online social network refers to a collective set of profile, content and network attributes and is defined as follows [10].

- *Profile attributes* describe their persona like username, name, age, date of birth, location.
- *Content attributes* refers to the content created or shared by the user such as the text, posts, tweets and the time of creation of the content.
- *Network attributes* refers to the network that a user makes inside an online social network like the numbers of friends or followers.
-

Identity resolution in social networks therefore refers to the process through which an identity is searched and analyzed between disparate identities to find a match and/or to resolve identities by using the above mentioned attributes of an identity.

The rest of the paper is organized as follows. In Section II we present the review of work done in identity resolution. In section III research challenges have been discussed. In section IV we propose a framework for identity resolution problem and discuss in length various phases of searching and linking the multiple identities of user scattered across online social networks. Finally we have concluded the paper in section V.

II. RELEATED WORK

The problem of matching identities in online social networks to check whether they belong to same individual has its roots closely related to fields like alias detection [3.16], author identification of

anonymous text [5, 23], matching the identities in blogs and discussion boards [1,7], identity deception [13, 14] and detection of imposters within social network [4, 8]. In this section we present the review of various studies that has been carried out to search and link the multiple identities of a single individual scattered across social networks.

By using supervised machine learning techniques [22] extracted 27 features of three types: name based, user information and social network topology based and constructed different classifiers to match users in two popular social networks: Facebook and Xing17. The proposed framework was evaluated on the real datasets from these networks and an accuracy rate of 98.2% was found in identifying the user profiles across two networks. Using the multiple sources of similarity like the similarity between the profile information, description of the user's interest and user's friends list, in [17] an approach was presented based on supervised classification for resolving user identities over social networks. The experimental were performed on data collected from Facebook, Flicker18 and LinkedIn and the results were compared using three different classification approaches (SVM, Random Forest and Alternating Decision Trees). The classification algorithms were found to be performing well in matching the identities, however SVM outperformed the other two.

In [11] a criminal identity resolution technique was proposed that examines three types of identity attributes: personal identity attribute, social behavior attributes and social relationship attributes. Relying only on the personal identity attribute didn't yielded better results. By incorporating the social behavior and social relationship attributes, significant improvement over the performance of matching technique was found.

Social network graphs have also been used for matching users across social networks. An unsupervised entity resolution algorithm was proposed in [12] that utilize both the attributes and connection graph of an entity for profile matching on online social networks. The algorithm works in a similar way as PageRank algorithm and circulates the similarity of each entity pair based on the connection graph. The attribute similarity of each entity pair is used to express that more similar entity pair in attribute should have larger PageRank. The algorithm was experimented on the profiles collected from Yahoo19 and Facebook and significant match was found. Moreover social graph are augmented with user attributes forming social attribute network and are not only used for link prediction or community detection but also for matching the accounts across social networks [15]. In [25] the social attribute network was generated for linking profiles where profile matching was based on social linkage and usage of profile attributes. This approach is suitable where profile data is poor, incomplete or hidden due to privacy concerns. Twitter and Facebook real datasets were used to evaluate the proposed algorithm. The results were satisfactory however, was not scalable to large graphs.

Stylometric features were also been used by researchers for the purpose of linking profiles across the social networks. Instead of the traditional approach of using profile information to match profiles across the social networks, only digital stylometry was used for matching the profiles on Twitter and Facebook networks in [29]. The model was evaluated on the 5,612 users and the correct match was found with the accuracy of 31%. On similar grounds rather than using the profile information, in [22] user's online shared material and social relationships was used for matching identities across Twitter, Facebook and LinkedIn by utilizing the string matching algorithms and natural language processing. The experiments show that the degree of accuracy increases with increase in the number of posts/tweets per user. Using five popular social network: Twitter, Facebook, Google+20, Myspace21 and Flicker [9] in their study explored large scale correlation of accounts across theses networks using various machine learning techniques for linking the accounts among these profiles based only on the user profile attributes (usernames, real names, cross names, location, photo, and face). Using the pair wise combination of Twitter, Facebook and Google+ the correct match was found to be with accuracy up to 80%.

Apart from the solution proposed by researcher community for matching identities across social networks, many online tools developed by commercial agencies are available. These tools can

locate an entity multiple references across the databases and online social networks. Some of the popular tools developed so far are *Pipl, Yasni, Peekyou, Spokeo*, *123poeple* and *OCEAN*. These tools simply work by taking an entity and their little attribute and search for the match across different databases and social networks. All these tools are commercial and do not provide any API or share data.

III. PROPOSED FRAMEWORK

The general framework for user identity resolution is given in in Fig-1. The framework has two major modules as feature extraction and model construction. In feature extraction module profile, content and network features are extracted from the profile of a user in an OSN. These extracted features then serve as an input to the model construction module wherein a supervised, semi-supervised or unsupervised model is trained on the labeled pair of profiles. The trained module is then used to predict the matched and unmatched identities of users across online social network.

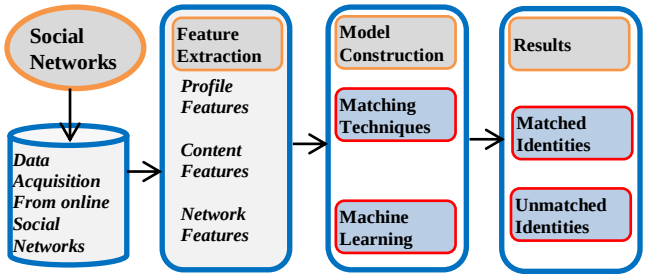

Figure 1. A framework for User Identity Resolution

A. Data Acquistion

The foremost thing to start with is acquire the data from OSN. This can be achieved in two ways; either is to use the existing datasets [18, 19] or generate your own dataset. The dataset can be built by using Application Program Interface (API) of OSN or by crawling OSN. Every OSN provides API for the developers to connect their applications with the network for data acquisition; however there are some restrictions and access limits imposed by the network providers when APIs are taken in consideration for data acquisition. Alternatives to APIs are to use the web crawlers and then scrap the content from HTML files of the network.

B. Feature Extraction

After data acquisition from OSN predefined profile, content or network attributes have to be extracted from the data so that they can be taken in consideration for the model construction. The attributes to be extracted vary from network to network, matching techniques uses and the model to be built. When Facebook and Twitter is taken in consideration for identity resolution we need to extract features from only those attributes which are common in both the networks so that we can find a match with high degree of confidence

C. Matching Techniques

To match the multiple identities of a user, several string matching techniques like Jaro-Winkler distance [30], Jaccard similarity [33], and Levenshtein (Edit) distance [28] are applied. Apart from string matching techniques Whiteprints and Timeprints can also be exploited when content posted by a user in an OSN is taken in consideration for user identity linkages. Writeprints also called as Stylometry is the statistical analysis of writing style where some specific characteristics or features of a piece of writing are analyzed to draw a conclusion on its authorship [23]. On the similar patterns the Timeprints of a user can also be created by extracting the publishing time of the contents posted by particular users. Time features can be used for unmasking the author identity in both the supervised (author identification) and unsupervised (alias matching or similarity detection) problems [7]. Social network graph of a user can also use for identity linkage across social networks where in a social graph can be generated from the friends and followers of a user.

D. Model Construction

By following the traditional way of data mining and classification, a model is to be built by exploiting the supervised [21,31], semi-supervised[2,31] or unsupervised [6,24] machine learning techniques. Supervised model is a typical example of binary classification problem wherein there are two types of

instances as matching identities (positive instances) and unmatched identities (negative instances). Classifier algorithms such as Na¨ıve Bayes, Decision tree, Logistic regression, KNN and SVM are employed in building the supervised model. Semi-supervised approaches takes in account both the labeled and unlabeled user identities wherein the unknown user identities are then predicted during the learning process. Unsupervised model can be employed in the unlabeled data where there is high cost associated with matching labeled identity pairs

IV. CHALLENGES

Although users with accounts on multiple social networks draw many opportunities however, drawing a link between the users identities scattered across social networks is not a straight forward task.

A. Limited Access:

To query the details of a particular user online social networks provide the Application Program Interface (API) for that. The private details are hard to get and can be made available only after convincing the user to share the private details which in itself is a challenging task. Public details are easily available however getting all the public details using API is also a challenging task as the number of public attributes retrieved by API varies from network to network. APIs of Facebook or Instagram shares only few public attributes while as API of twitter shares most details of a user

B. Credibiity Of An Account

Social networks allow a user to create an account for free or at a very low cost. It is challenging to distinguish the trustworthy from untrustworthy users. There have been many evidences found where a fake account is being treated as a legitimate account in Facebook or Twitter and the friends and follower are more than that of legitimate accounts. Heterogeneous nature of Social Networks

C. Hetegernous nature of Social Network

To avail the services offered by social network a user registers on a social network by defining the profile attributes during the registration so that the user is uniquely identified as well as helps the others in finding them. The quantity and granularity of information varies within each online social network, with some demanding descriptive attributes while few need a valid email and chosen username [10]. Hence the attributes that will be put in use for identity resolution in a particular social network might fail in the other network.

D. Attribute Evoultion

A user on social network defines his profile attributes while creating an identity in social network. With the passage of time he continues to update his attribute. It may be the case that a user updates his attribute only in particular social network which he uses the most. In these cases it may lead to the inconsistency when the identity linkage is primarily on profile attributes.

V. CONCLUSION

Identity resolution is of great importance in many application areas such as recommendations, link prediction, de-duplicating audience, security practitioners etc. This paper introduced a general framework for identity resolution problem wherein the various phases of the framework were discussed in length. The state of art work previously done by academicians in identity resolution problem and some closely related fields were also discussed. Moreover some challenges that may be encountered tackling the user identity linkage problem was also presented in this paper. As a future work we will be implementing this framework on two popular online social networks Facebook and Twitter.

VI. REFERENCES

[1] Amrendra, S, “Visualization and Detection of Multiple Aliases in Social Media”. Master’s Thesis, Upsala University, Sweden, 2013

[2] Arvind N., and Vitaly S., “Deanonymizing social networks.”, In ISSP, 2009

[3] Arvind. N., Hristo, P., Neil, Z., G., and John, B.,“On the feasibility of internet-scale author identification,” in 2012 IEEE Symposium on Security and privacy (SP), 2012, pp. 300 –314

[4] Bilgic, M., Licamele, L., Getoor, L., and Shneiderman. B., “D-dupe: An interactive tool for entity resolution in social networks”, In2006 IEEE Symposium on Visual Analytics Science and Technology, 2006

[5] Brocardo, M. L., Traore, I., Saad, S. and Woungan, I., "Authorship verification for short messages using stylometry". 2013 International Conference on Computer, Information and Telecommunication Systems (CITS), Athens, 2013, pp. 1-6.

[6] Christopher R., Yunsung K., Augustin C., Nitish K., and Silvio L., "Linking users across domains with location data: Theory and validation". In WWW, 2016

[7] Fedrick, J., Lisa, K., and Amrendra, S., "Time Profiles for Identifying Users in Online Environments," 2014 IEEE Joint Intelligence and Security Informatics Conference, The Hague, 2014, pp. 83-90.

[8] Fong, S., Zhuang, Y., and J. He, "Not every friend on a social network can be trusted: Classifying imposters using decision trees," The First International Conference on Future Generation Communication Technologies, London,2012,pp.58-6

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

[10] Jain, P., "Automated Methods for Identity Resolution across Online Social Networks". Doctorate Thesis, Indraprastha Institute of Information Technology Delhi, 2016

[11] Jiexun, L., and Alan, W., "Criminal identity resolution using social behavior and relationship attributes," Proceedings of 2011 IEEE International Conference on Intelligence and Security Informatics, Beijing, 2016, pp. 173-175

[12] Limgfenu, Niu, Jiamin, W., and Yong, S.,"Entity Resolution with Attribute and Connection Graph," 2011 IEEE 11th International Conference on Data Mining Workshops, Vancouver, BC, 2011, pp. 267-271

[13] Michail, T., "Identity Deception Prevention Using Common Contribution Network Data," in IEEE Transactions on Information Forensics and Security, vol. 12, no. 1, Jan. 2015, pp. 188-199

[14] Michail, T., and Sherali, Z., "Multiple Account Identity Deception Detection in Social Media Using Nonverbal Behavior," in IEEE Transactions on Information Forensics and Security, vol. 9, no. 8, Aug. 2014, pp. 1311-1321,

[15] Neil, Z., G, Wenchang, X., Ling H., Prateek, M., Emil S., Vyas, S., and Dawn S.,"Evolution of social-attribute networks: Measurements modeling, and implications using google+". In Proceedings of the 2012 ACM Conference on Internet Measurement Conference, IMC '12, 2012,pp. 131–144

[16] Novak J., Raghavan P., and Tomkins A., "Anti-aliasing on the web," in Proceedings of the 13th international conference on World Wide Web. New York, NY, USA: ACM, ,2004, pp. 30–39

[17] Nunes, A., Calado, P., & Martins, B. "Resolving User Identities over Social Networks Through Supervised Learning and Rich Similarity Features". In Proceedings of the 27th Annual ACM Symposium on Applied Computing, 2012, pp. 728–729.

[18] Oana G., Howard L., Sree H., Gerald F., Robin S., and Renata F. "Exploiting innocuous activity for correlating users across sites", In WWW, 2013

[19] Oana G., Patrick L., Robin S., Renata T., and Krishna P G., "On the reliability of profile matching across large online social networks". In KDD, 2015

[20] Pew Research Center, Social Media Matrix. http://www.pewinternet.org/2015/01/09/social-media-update- 014/pi_2015- 01-09_social-media_10/ [Online; accessed 18-March-2018]

[21] Reza Z., and Huan L., Connecting users across social media sites: a behavioral-modeling approach, In KDD, 2013

[22] Reza, S. and Abdolreza, A., "Identity matching in social media platforms," 2013 International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS), Toronto, ON, 2013, 2013, pp. 64-7

[23] Rong, Z., Jiexun, L., Hsinchun, C., and Zan, H.,"A framework for authorship identification of online messages: Writing-style features and classification techniques". J. Am. Soc. Inf. Sci. Technol., 57(3):378 - 393, ISSN 1532-2882, 2006

[24] Lebastian L., Irina T., and Hannes H.'" What your friends tell others about you: Low cost linkability of social network profiles".In KDD, 2011

[25] Sergey B., Anton K., Seung P., Wonho R., and HyungdongLee., "Joint Link-attribute user identity resolution in online social networks". In SNAKDD Workshop, 2012

[26] Soryani, and Minaei, B., "Social Networks Research Aspects: A Vast and Fast Survey Focused on the Issue of Privacy in Social Network Sites", arXiv preprint arXiv:1201.3745, 2012

[27] Stastista, Most famous social network sites worldwide as of January 2018.

https://www.statista.com/statistics/272014/global-social-networks-ranked by-number-of-Users/ [Online; accessed 18-March-2018]

[28] Tereza I., Peter F., Fabian A., and Kerstin B., "Identifying users across social tagging systems", In ICWSM, 2011

[29] Vosoughi, S., Zhou, H., and Roy, D.,"Digital Stylometry: Linking Profiles Across Social Networks". Social Informatics: 7th International Conference, SocInfo 2015, Beijing, China, December 9-12, 2015, Proceedings pp. 164–177.inbook, Cham: Springer International Publishin

[30] William C., Pradeep R., and Stephen F., "A comparison of string metrics for matching names and records".In KDD, 2003

[31] Xiaoping Z., Xun L., Haiyan Z., and Yuefege M., "Cross-platform identification of anonymous identical users in multiple social media networks". IN TKDE, 2016

[32] Xin M., Feida Z., Zhi-Hua Z., and Jianzong W., "User identity linkage by latent user space modeling" In KDD, 2013

[33] Yilin S., and Hongxia J., "Controllable information sharing for user accounts linkage across multiple online social networks". In CIKM, 2014