Multi-attribute identity resolution for online social network

Shalini Yadav¹ · Adwitiya Sinha¹ · Pawan Kumar²

Received: 28 August 2019 / Accepted: 14 November 2019 / Published online: 21 November 2019
© Springer Nature Switzerland AG 2019

Abstract

Social media has gained prominent immensity in usage with the explosion of communication owing to social phenomena. An apparent dimension of social media is the virality of information, which when generated from redundant and false user identities, may cause chaos across online social communities. Duplicate user profiles are quite conventional in social networks causing unintentional faults or intentional deceptions. Such fraudulent motives may involve cyberbullying, fake boosting, fake following, cyberstalking, etc. Hence, revealing the malicious users having multiple profiles in social networks has become significant for downstream analysis to ensure cyber safety. Identity resolution is considered one of the pivotal techniques to reveal redundant identities, which are found co-referent to the same real-world user. As per social media theory, there are different stages for automated identity resolution, namely identity searching, identity linking and identity merging. Our research is focused on developing a novel identity resolution framework for detecting redundant user profiles on Twitter social media constructed with nodes, ranging from small-scale to massive-scale networks. Our proposed solution extracts Twitter user profiles with various attributes, for instance, first name, last name, username, user id, tag line, location, language, profile URL and tweets. We developed various algorithms for matching and merging redundant user profiles with the Jaro–Winkler similarity technique. The similar profiles are linked to eradicating their distinctive impact from the originally constructed social network. Our experimental outcomes illustrate iteration-wise reduction of irrelevant and redundant profiles for the close-community of the randomly selected user, which is further extended to the entire user-based and trend-based Twitter network. The results of our approach were compared with existing counterparts and were found to excel in performance concerning accuracy in the detection of redundant identities. Our approach would greatly assist viral marketing, terrorist screening, and social media trending.

Keywords Social media · Identity resolution · Fake profiles · Identity merging · Social influence · Viral marketing

1 Introduction

Social networking is an emerging field in applied sciences, which employs existing scientific knowledge & computational paradigm to develop practical solutions for societal applications. Social networking assists in understanding the social interactions among individual social entities, social groups, and online organizations. Social interaction can be used to describe friendship, connections, followers, etc. Generally, social networks are in spontaneous order,

efflorescent and complex. The global logical patterns in interactions appear from the local acquaintance of the users who build up the network. As network size increases, these patterns become more superficial. There are several Social Networking Services (SNS), which offer an online platform that could be used by people to engage in social dialogues. Some of the prevalent instances include MySpace, LinkedIn, Twitter, Facebook, etc. People can share their personal or course interests, activities, likes movies, books, etc. The lifeblood of SNS is User-Generated Content

✉ Adwitiya Sinha, mailto:adwitiya@gmail.com; Shalini Yadav, yadav.shalini015@gmail.com; Pawan Kumar, pawankumar.063@gmail.com
| ¹Computer Science Department, Jaypee Institute of Information Technology, Noida, Uttar Pradesh, India. ²ALISDA, DGAQA, Ministry of Defence, Bengaluru, India.

(UGC) such as text posts, comments, tag photos, etc. Most of the social networking services are web-based and provide a platform where the users interact with e-mails, messaging, etc. Social networking services are used to create a network of people having friends and followers. The significance of social networking sites is all about sustaining personal and business relations with the on-going trend. All SNS is becoming a powerful marketing and communication tool.

With various emerging social platforms, certain significant challenges become inevitable, especially to manage fake and redundant profiles. This demands for Identity Resolution (IR). It is referred to as a process that allows an identity to be searched and analyzed among the pool of social identities and relational identities [1–3]. For instance, on Twitter, an individual social profile is associated with different parameters, such as tweet activity, comments, posts, followers, impressions, etc. IR focuses on such attributes to matching profiles to reveal the redundant profiles. The following section highlights stages through which identity resolution is performed for managing redundant social profiles on the social platforms.

- **Searching** This involves processing the list of user identities on the social network, which is similar to an existing profile [1]. It is to leverage the set of complete attributes namely, profile search, content search, self-mention search, and network search. Profile search gives us basic information about the user's identities. Content search where people create or share with others. Self-mention searches the method to exploits the user's tendency and to cross-pollinate the information on the online social network. For searching the candidate identities, profile searching is normally employed to extract basic information, usually provided by the user while creating his profile.
- **Linking** This procedure computes the similarity between identities obtained from the identity searching list [1]. The methods include syntactic linking, semantic linking, crowd-source linking, graph linking, behavioral linking, and image linking. Syntactic and semantic methods calculate the similarity by some different techniques involving edit distance, Jaccard distance, Jaro distance, Soundex, ontology matching, etc. However, it only works on the string-based profile or content attributes. Image linking calculates the similarity between the two profiles which is used by the two identities. Graph linking derives the friend connections network similarity of the two identities based on their activity and interaction. Crowd-source linking generates the human perception to affiliate matching score of each user identity based on the user's knowledge.

Behavioral linking methods are mostly based on the username creation and username reuse.

- **Merging** This stage is performed for matching candidate identity, obtained from identity linking. Based on the similarity matrix, attributes values of user identity are matched. The set of matched identities are merged to reveal unique user identities, thereby removing the ambiguity of user accounts and online profiles.

Users may create multiple accounts for various reasons, which may include research, personal use, promotion, or malicious motives (fake boosting, doxing, online stalking, etc.). Finding the unique users in a network is a challenging task, hence necessitates IR methods on available users profiles. This allows multi-attribute analysis based on a set of attributes on the user profile. Reducing the graph with merged links and user accounts results in a more authenticate and reduced network, free from multiple repeated accounts.

1.1 Significance of an identity

The identity of the user in a social network is self-assured in three attributes that include personal, social, and relational (Table 1). Personal attributes refer to the ones that usually include personal information. Social attributes append those items, which are exchanged with the social network platform being shared by users in the online social network. Relational attributes are correlated attributes that link online friends and connections [2].

1.2 Mathematical foundations of identity resolution

For users with identity I_A obtained from an online social networking site, signified by SN_A . The following expression denotes the correct identity, I_B on other online social networking sites, SN_B .

$$I_A \rightarrow I_B \quad (1)$$

Table 1 Three identities and associated attributes

Class	Attributes
Profile attributes	First name, last name, gender, location, education, profession, email, language, date of birth
Content attributes	Tweets, video posts, image posts, YouTube links
Network attributes	Friendships, group membership, fan page participations, connections, followings, and followers

Identity Resolution follows two sub-processes in online social network namely, identity search and identity linking. The process of identity searching is to find the candidate identities set of lists on SN_B which is similar to a given identity I_A . The identity linking process is to calculate the similarity matrix between identity I_A and all candidates which returned by identity search procedure. This is followed by candidate identity ranking based on the similarity matrix and identities which have matched score is returned as, I_B .

a. Identity searching process

Given user identity I_A and search attribute S , the problem is to find a set of candidate identities, I_{Bj} on a social network, SN_B forming relation $S(I_A) \sim S(I_B)$ [2].

$$\{I_A, S\} \rightarrow \{I_{B1}, \dots, I_{Bj}, \dots, I_{Bn}\} \quad (2)$$

Search attributes can be found on the identity search method, including *Profile*, *Content*, *Network*, and *Self-Mention* search. Profile search implies the search of candidate identities on SN_B by the search attribute fetched from identity I_A . Content Search implies the search of candidate identities on SN_B with the attributes of the content of I_A as search attributes. Network Search implies the search of candidate identities on SN_B with the attributes of the network of I_A as search attributes. Self-mention search implies to explore those identities I_A which created their content in the form of URLs or post URLs in tweets. Pointing to one's own identity I_A , is referred to as self-mention.

b. Identity linking process

Given user identity I_A on social network SN_A , a candidate identities set $Q = \{I_{B1}, \dots, I_{Bj}, \dots, I_{Bn}\}$ on social network SN_B and match function M , maximum of identity pair (I_A, I_{Bj}) with highest similarity match score I_{Bj} implied as I_B [2].

$$\{I_A, Q, M\} \rightarrow \{I_A, I_{Bj}\} \rightarrow I_B \quad (3)$$

Identity linking algorithm is built to calculate the similarity between the identities and network of candidate identities set which is obtained from identity search concerning profile attributes. A high similarity value among a set of profiles was matched and linked. Table 2 highlights

the data fields, which can be considered for experimentation with different social networking sites [3].

Our initiative is built for Twitter-based users, hence semantic and syntactic methods are used. These methods are usually applied for the profile-based attribute having string-based value, hence it is used for our Twitter case study. The similarity is derived using Jaro–Winkler distance-based similarity technique. It compares the strings and generates the similarity score. The pairs having a similarity value greater than the threshold are combined in the cluster. Our study also applies correlation, for estimating better accuracy in measuring the similarity between multiple trends.

c. Identity merging process

Given Cluster list L , obtained from social network SN_A and SN_B similarity matrix of their attributes $L = \{c_1, \dots, c_n\}$. Target $T = \{a_1, \dots, a_n\}$ is set track profile matches.

$$T_i \cap L_i = T_i \quad (4)$$

After, identity searching and identity linking similar pairs of candidate identities are revealed having the highest similarity score and correlation. Through string matching, a combination of two nodes resulted in a single node where all attributes values form the union of both the nodes. Through this, single individual profile holders in a social network are revealed.

2 Related works

Several significant research has been conducted for improving the methodology of revealing redundant and fake user profiles on the social network. The authors in [4] introduce matching of profiles, in several scenarios such as data integration, data enrichment, information retrieval, etc. They hypothesize profile information as homepage, username, location, etc. Two profiles describing the same physical person whose proprietorship values are the same also called Inverse functional Property (IFP). Aggregation method is used for data fusion and decision making to discover the maximum number of profiles that refer to the same person between two online social networks. Three

Table 2 Profile data extracted from a social networking platform

Social networking site	Parameters
Twitter	First name, last name, username, id, tag line, location, language, profile URL, tweets
Facebook	First name, last name, username, date of birth, location, id, gender, about, language, email, website, occupation, education, text/video/image posts, profile URL

main areas are inspected such as social network profile heterogeneity, the similarity between attribute values, and deciding whether two profiles are the same or not. In [5] authors focussed on the rapid growth of popularity through online social networking services. Information on user identities was extracted from different social networking sites. The context-specific technique is used to measure the similarity from different social networking sites mainly for Twitter and LinkedIn networks.

Research conducted by authors in [6] proposed a graphlet method for entity resolution to find relationships between identities and using these relationships to map identities to the available entities. Graphlets are collections of small graphs that are used to describe the role of a node in a graph and provide features to describe the authenticity of users. The authors used objective function, attribute fit to capture the same identity which is available as two separate profiles. Another function, the relational fit was used to capture a single user who interacts on two different social media sites. In [7] people use various social networking sites for different purposes. When information from different sources integrated, then combined social data of users will have a broad prospect in the area of recommendation system and security. The key problem from combining social data is to discover multiple accounts of a person across different social networks. The author proposed a method for user identity resolution based on personal profile information and their social network. This approach is completed by calculating the similarity between two vertices. Using the Kleinberg's distance, to measure the similarity between two graphs that produce a similarity matrix and compare relations between two accounts.

Automated methods for identity resolution on social network accounts were proposed in [1]. To address the identity resolution process, the authors applied similarity techniques to search for data individually and combine the profiles. Further, research conducted in [8] authenticates genuine against fake accounts. The supervised machine learning algorithm is applied to build a knowledge base and features. In [9] authors described identity resolution as the potential use for social networking profiles which is relevant to the various purpose. It forms open-source intellect applications to construct a semantic web network. Representation of research in the social network is inhibited by the lack of data. It links the identity of data sources which previously used by the researchers that were no longer available. They evaluate the method for an identity resolution with an easy approach to a realistic labeled dataset of online social profiles and draw the largest and significant online social network. It finds the primary network which concedes more rapid collection and repetition of profiles. It would benefit from improving

the collection rate of goals for reproducibility in identity resolution research. In [10] authors describe social networks that cover a wide range of the population. According to a privacy level, people make multiple networks to enjoy services and sharing personal information. In social networks like Twitter, Instagram, and Weibo, social links represent the interests rather than the persons that they know. It proposes a community detection method based on the interests group and applies the de-anonymization algorithm. The de-anonymization algorithm was used to attack social network privacy and conducting a large graph. Community detection methods mainly based on the graph structure. It also has strong and weak community detection methods.

Joining social networks leads to the creation of identity across three major dimensions: profile, content, and network [2]. Users' online social identities are often unlinked, isolated and difficult to search. It proposed searching techniques based on profiles, but it leaves another identity such as content and network as unexplored. It introduces two identity search methods based on content and network attributes and improving the identity search algorithm based on the profile attributes of the user. This leads to the application of the proposed identity search algorithm for finding a matched user identity on Facebook, Twitter, and LinkedIn. In [11] online social network is a great place for fake people to portray the identities of people via creating fake profiles. Fraudulent people carry out malicious activities such as harming user's prestige, isolation and random attacks on online social network sites. Finding out the fake profiles accounts is one of the toughest security problems in the online social network. Fake Profile Recognizer (FPR) is a technique for observing and deleting fake profiles in social networks. The disclosure methodology in FPR is based on the technique of regular expression (RE) and deterministic finite automaton (DFA) for observing identities of profiles.

Research in [3] highlights about importance of identity in profile data integration, identity resolution, etc. and finding similar profiles of an individual user among Facebook, Twitter, etc. It only considers profile attributes such as name, location and treats users solely as a medium for online shared material and social relationships. String matching algorithm, natural language processing, and API evaluate and compare profiles. Profiles are considered as a match if their similarity value is above the predefined threshold. In [12] authors introduce expeditious growth of smartphones and novel advancement in telephony. Some users have more than one calling identities for several purposes such as business, personal, etc. making multiple identities for a single user. It is compulsory to delete fraud identities, criminals and analyzing the total social network graph of an individual. For finding the identity of a single

user, they describe the ID-CONNECT approach for multiple identities which is based on a single social call graph and call behavior. Finding similarity between two identities, it can be better predicted by using both call network and behaviour of user identities towards their connections. The higher similarity between graphs of two user identities is a two-step approach i.e. predicting similarity by call rate and time duration, generating a candidate set of identities for each user identity.

Personal profiles of people generally contain delicate information, which may occasionally get leaked and disturb the people's privacy [13]. A user's self-controllable matching profile protocol in preserving the privacy of the mobile social network is being carried out in this regard. The protocol allowed users to design their matching metrics to include their matching options and making the matching result more stringent. Detail analysis includes protocol finding, privacy protection of both, profile attributes and attributes values during the matching procedure. Extended evaluations are managed to clarify that their protocols are more efficient rather than available protocol with respect to computation and communication principally when the profile attribute value is large. Protocols use a weighted Manhattan distance-based similarity technique in which weights and the threshold value are chosen by the user itself. In [14] authors describe that users can be analyzed across social web tagging systems by combining user id and tag. Different approaches are introduced for comparing and measuring the distance between profiles for identification. Users are identified in the different services on the basis of their social tags. It also investigates the deliberation of network structure in a combination of tag-based profile users. In [15] authors describe that presently, it is common for users to have redundant profiles on different social networks. While discovering multiple accounts of a person across the different social networks that allow us to merge all social connections from different services and making a complete graph which is helpful in many applications. The author proposed an approach for matching user account by conditional random field. It is appropriate when data is lossy, incomplete and hidden from others, owing to privacy settings. It also focuses on the social links for an identity resolution that shows profiles are matched on the basis of social relationships existent among online users.

The significance of the topological structure of the network in detecting community, irrespective of content analysis, plays an eminent role in identity resolution [16]. Each cluster in the community have the same topic of interest, hence a relation was built among clusters by social objects to link and resolve similar identities. Social media is a new communication medium for information retrieval and propagation that connects people with similar interests

[17]. The research, conducted in [18], authors describe the detection of dubious links in the user community based on mutual cluster coefficient and user-centric account information. Fake users in the community are exploited by mutual friend features. It also identifies strong and weak ties of users.

In [19] authors describe exponential growth of web-based online social networks in the past few years. A user on an online social network is distinguished mainly by three attributes such as profile, content, and network. In [20] a person has multiple accounts on online social media to interact with other persons. The user-generated post is used to match the identity across the social network. In the recent survey, almost one-third of persons in the world have a profile on social media. The data of the same user on different social media networks is necessary for integrating the result. A user generally posts identical data publicly on different social media among multiple users to get popularity. In [21] due to the increase in electronic storage, false and duplicate records are quite often in nature. Identity resolution is based on the similarity method of various attributes and algorithms. The commonly available attributes such as name, gender, date of birth and identification number are primarily used for identity resolution. Biometric features are considered a more reliable attribute for identity resolution. In [22] authors describes the significant impact of user profile attributes on identity resolution in the social network. Most of the users register on the social network according to their usage. If more common data is available then there are higher chances of accuracy in matching the social profile of users across the social network. In [23] authors describe the importance of profile linking in the social network that helps in marketing strategy, friends suggestion, a platform for discussion, etc. Most of the users have the same profile name across the various social network during registration. Similarity among user profiles is calculated on the basis of various features. The activity of a user can be observed in the other platform by *Link Social*. Linking several social media platforms help in understanding user online behavior and other actions. In [24] authors illustrate real-time information in the Twitter social network on the basis of various network parameters. In [25] authors describe the profile linkage of people having the same identity. In today's scenario, people connect to their friends via social networks. The probabilistic classifier method links real-world large scale profiles. People have different user names across the different social networks so that the linking of user-profiles becomes complex. Further in [26], authors describe public and private information available on social media. Profile information is used to classify users across social media. In [27] entity resolution matches profiles from the different social networks. It is assumed that most people have more

than one account on social media. The similarity score is used to compare online user profiles from different social networks. Further, in [28] authors describe a similarity-based link prediction method that offers future implications of the existence of any new connections among different users or groups on the social network that are likely to evolve.

On a social network, people can find or share new information which is popular at that time. Identity resolution helps in recommendation system, fraud detection, profile integration, finding and associating identities of an individual across the social network. Motivated from the above research in the related area, we propose to design a novel technique to detect redundant profiles on Twitter algorithmically and merge them, to minimize the multiplicative impact of the existence of such fake and irrelevant profiles on social media.

3 Proposed design and modelling

Our proposed model is developed with the following research goals that can be represented as significant stages as depicted in Fig. 1.

- **Attribute selection** Our proposed system collects user attributes from Twitter profiles including, first name, last name, user handle, id, tag line, location, language preference, profile web link and tweeted text. The data collection phase is preceded by the construction of a social network graph with the following and follower relationship, which is extended up to two levels of association.
- **Similarity matrix** Algorithms for matching and merging of redundant user profiles are conducted with the Jaro–Winkler similarity technique. This stage facilitates the formation of a matrix that matches similar users' identities in terms of all the attributes taken.
- **Cluster formation** Similar user identities are further clustered as a single group. The clustering of profiles is carried out based on high similarity value, preferably greater than the predetermined threshold.
- **Correlation computation** This stage computes the correlation score of the user accounts in terms of the profile attributes. Our collective approach detects the user

profiles that are co-referent with the same real-world individual profile.

- **Profile merging** The multiple correlated profiles clustered as a single entity is further merged. Finally, the network with reduced nodes is plotted, after the merging of user identities. Initially, the experimental outcomes illustrate iteration-wise reduction of redundant profiles for a close-community of the randomly selected user, which is later extended to the entire user-based as well as trend-based Twitter network.

Our proposed framework for the detection of redundant user profiles on the Twitter social network is detailed in Fig. 2. Initially, Twitter user attributes were crawled from using its associated Twitter API and account extractor. All the user records are buffered in the database. This is followed by processing the stored user profiles with selective attributes to apply the Jaro–Winkler similarity technique, resulting in similarity matrix formation that is further used for correlation computation. Finally, matching is performed against all the user identities, using our algorithms. Resultantly, the highly correlated user profiles, clustered as a group, are merged to eradicate the redundant impact of the repeated accounts over the network. Our proposed algorithms are executed and results are compared with existing benchmark approaches. The accuracy and scalability achieved in the detection of redundant identities using our framework were more than the referenced counterparts. The mathematical expression for similarity computation is further elaborated in the following section.

Jaro–Winkler is applied to compute the similarity between attributes for another attribute on the basis of string-based value. The Jaro distance a_j of two given strings b_1 and b_2 can be formulated as in Eq. (5):

$$a_j = \begin{cases} 0 & \text{if } m = 0 \\ \frac{1}{3} \left(\frac{m}{|b_1|} + \frac{m}{|b_2|} + \frac{m-t}{m} \right) & \text{otherwise} \end{cases} \quad (5)$$

where social parameter $|b_i|$ denotes the length of string b , m is the number of matching characters and t refers to half of the total number of transpositions. Two characters from b_1 and b_2 respectively considered for matching only if they are the same or not $\frac{\max(|b_1|, |b_2|)}{2} - 1$. Distance metric uses prefix scale p on the string which gives rating and match string from starting of a prefix length. Given two strings b_1 and b_2 and Jaro–Winkler distance a_w .

Fig. 1 Significance stages in proposed methodical approach

Fig. 2 Proposed framework design

$$a_w = a_j + (l_p(1 - a_j)) \quad (6)$$

In Eq. (6), parameter a_j corresponds to Jaro distance computed in Eq. (5). l is the length of the common prefix which marks the beginning of the strings and p is a constant factor. The threshold value in Jaro–Winkler’s implementation is set to 0.7.

$$a_w = \begin{cases} a_j & \text{if } a_j < b_t \\ a_j + (l_p(1 - a_j)) & \text{otherwise} \end{cases} \quad (7)$$

The hierarchical cluster data mining technique is a method of clustering analysis that gives a hierarchy of clusters. It is classified into two types, agglomerative and

divisive clustering. In agglomerative clustering, observations start in its clusters and cluster pairs are merged when it moves up in the hierarchy (bottom-up approach). Divisive clustering is a top-down approach, where all observations start in the network and split them down the hierarchy. The complexity of agglomerative clustering is $O(n^2 \log(n))$, while that of divisive clustering is $O(2^n)$. Hence, owing to reduced complexity, agglomerative clustering is applied to our experimentation with Single-LINKage clustering (SLINK) and Complete-LINKage clustering (CLINK).

The maximum distance between the similar attributes clusters becomes $\max\{d(x, y) : x \in A, y \in B\}$ and the

minimum distance between the similar attributes clusters (SLINK) are given by $\min\{d(x, y) : x \in A, y \in B\}$, where x, y are entities belonging to a cluster set A and B , respectively. Finally, the mean distance between the similar attributes cluster becomes $\frac{1}{|A| \cdot |B|} \sum_{x \in A} \sum_{y \in B} d(x, y)$. The identified redundant nodes are merged by string matching technique, in which nodes with similar attributes are considered a single node. The following section describes the proposed framework involving detecting and resolving the impact of redundant and irrelevant user identities.

4 Proposed implementation framework

Our research contributes to identifying the redundant profiles having multiple accounts being managed by a single master profile for the various malicious purpose involving, cyberbullying, fake profile boosting, online shamming, doxing, etc. In the dataset, our focus was on personal connections in a social network which differs in strength. Such connections among people are often sparse and weak. The possibility of a strong social network occurs within close groups and friend circles. However, weak network ties occur over bigger communities at large. Our dataset is built over users profiles on the Twitter social network with 643 nodes and 792 edges (large-scale). Initially for checking algorithm, we take a single user network which has 170 nodes and 197 edges (small-scale). In this dataset, the friendship network is constructed with the Twitter user (@sjaswinder300) and his top 5 followings and followers

at the first level and the next level is constructed with followings and followers of users extracted the list of following and followers. Further, a trend-centric graph Twitter graph is also considered to apply and verify our proposed algorithms for a massive-scale Twitter network with 10,000 nodes.

Twitter is an open social network platform where people may read and post short comments in 280 characters of tweets. People usually follow twitter for information and entertainment purpose. According to research conducted in April 2014, 44% of user accounts have never tweeted, which means that there exists a considerable number of fake profiles that are possibly managed by a single master profile. Therefore, we focussed on the Twitter social network to perform our case study. Data extraction from the Application Programming Interface (API) of Twitter requires a handshaking mechanism, which is diagrammatically highlighted in Fig. 3 and its major constituents are elaborated as follows:

- *Resources owner (user)* It is the user who authorizes an application to get access to their accounts. Application access to the user account has limited scope.
- *Resources server and authorization server (API)* Resource server protects all the user account that exists on Twitter. The Authorization Server verifies the user's identity and provides access to the application.
- *Client (application)* The client is the person who may access user accounts through authorization by API.

Fig. 3 Twitter authorization process

Our experimentation is carried out in RStudio with associated packages and APIs.

4.1 Proposed identity search process

In our proposed model, identity searching begins with the extraction of followers and followings of the user, henceforth fixed as a seed node. This is followed by the extraction of attributes of followings and followers, containing a name, screen name, language, and location. For the Identity search process, some libraries are used from RStudio. TwitterR is an API to get access to the Twitter account and the RCurl library is used as a client interface for R that acts as a wrapper. It provides a function to allow the user to make general HTTP requests and also provide a general function to fetch URIs, post forms, etc. and the result is returned by Web Server.

4.2 Proposed identity linking process

Our proposed profile linking approach was carried out on the set of searched users (from the previous stage) with reference to the set of similar attributes. The string distance matrix computation technique was used on attributes for finding similar profiles. Further, correlation is used for measuring the accuracy of the computed similarity value. If the similarity measure between a pair of the profile is greater than the threshold based on user attributes, then the profile pair is termed similar. This procedure is illustrated in the following Algorithm 2.

Algorithm 1. Proposed Identity Searching for Social Network Creation

Input:

CusKey : Consumer Key
CusSecret : Consumer Secret
AccKey : Access Key
AccToken : Access Secret Token

Procedure:

```

1. Begin
2. setup_twitter_oauth (CusKey, CusSecret, AccKey, AccToken)
3. getUser = Id // give one user Id
4. lookupUsers(getFollowers = Id) // ids of followers
5. sapply(followers name) // followers name
6. lookupUsers(getFriends = Id) // ids of followings
7. sapply(friends name) // followings name
8. DF1(Followers, Following) // data frame creation
9. for each (Id in getFollowers)
10. getUser = Id
11. lookupUsers(getFollowers = Id)
12. sapply (followers name)
13. lookupUsers(getFriends = Id)
14. sapply(friends name)
15. DF2 (Followers, Following) // creation of second data frame for lookup
16. followers = rbind (DF1, DF2) // combine two data frames
17. end for
18. rbind(followers, following)
19. SN ← tkplot (DF) // plotting data frame as social network graph
20. end

```

Output: Followings & Followers Social Graph, *DF*

Algorithm 2. Proposed Identity Linking**Input:**Users Profile Attributes [i] = name, screen name, location

Threshold: similarity threshold value is 0.7

Similarity (S): similarity value between 0 & 1**Procedure:**

1. Begin
2. For each attribute //single attribute
3. similarity (S) on attribute // similarity matrix
4. if (similarity matrix of attribute > threshold)
5. Clustering on the similar attributes
6. else
7. neglect those similar pairs
8. end if
9. $k = 1 \dots (n - 1)$
10. $X = \text{dataframe}(\text{attribute}, \text{cutree}(\text{clusters}, k))$
11. Create list of clusters
12. kable() // output as table format
13. for each [cluster pairs in list]
14. Calculate correlation_matrix
15. Plot correlation_matrix // correlation graph
16. end for
17. attribute += attribute[i] // union of all attributes values
18. end for

Output: Similarity Matrix, Clusters Formation, Dendograms, Correlation_Matrix, Correlation Graph

4.3 Proposed identity merging process

The proposed identity merging procedure is illustrated in Algorithm 3, which joins all those linked profiles

(from the previous stage) that were found belonging to a particular node, represented as the respective master profile.

Fig. 4 Social connections fetched from twitter network forming **a** small-scale and **b** large-scale network

Algorithm 3. Proposed Identity Merging**Input:***list*: list of clustersSet $L = 0, K = 0$ Set n, m **Procedure:**

1. begin

2. Remove Punctuation (*list*)3. Sorting(*list*)4. for each *list* of row *i*5. for each *list* of col *j*6. $n = \text{list}[i][j].\text{length}$ 7. $m = \text{list}[i][j + 1].\text{length}$ 8. while ($L < m$)9. if($\text{list}[i][j].\text{char}(K) == \text{list}[i][j + 1].\text{char}(L)$)10. $K++$; $L++$

11. else

12. $L++$; $K = 0$ 13. if ($K == n - 1$)14. return $\text{list}[i][j]$

15. end while

16. end for

17. end for

18. plot $\text{list}[i][j]$ 19. $M_{SN} \leftarrow \text{list}$ **Output:** Merged Graph, M_{SN}

5 Experimental outcomes

This section highlighted the results from experimentation carried out over user-centric and trend-centric graphs. User-centric graphs include single-user Small-Scale Network (SSN) and multi-user Large-Scale Network (LSN); while the trend-centric graph is constituted as Massive Twitter Network (MTN). Figure 4a represents SSN with a blue node denoting the seed user and grey nodes

representing his direct associations, i.e. followings and followers. Figure 4b highlights the large-scale network formed with the list of followings and followers (grey nodes) of the seed user's direct associations (red nodes). Hence, it provides a pictorial layout of two-hop neighborhood connectivity, thereby forming a large-scale network.

Figure 5 shows the experimental outcome of our proposed identity resolution algorithm being executed on single-user SSN with redundant users being highlighted in red. In Fig. 5a, it is apparent that the revealed number of redundant profiles are initially quite high. However, the number of redundant profiles decreased to 166 users with 191 links after merging identified redundant users, as shown in Fig. 5b. Finally, the proposed model is executed on the multi-user Large-Scale Network (LSN) for extended analysis which is shown in Fig. 6. Initially, LSN has a higher number of redundant profiles (highlighted in red) as apparent from Fig. 6a. After all similar pairs of user-profiles are detected and merged as the union of their links, the number of nodes finally got reduced to 638 nodes bearing 785 edges as evident from Fig. 6b.

Figure 7a, b shows the decrease in the number of overall authenticate users in single-user SSN and multi-user LSN, which is a clear indication of the successful merging of redundant profiles on the Twitter social network. Further, a comparative study is conducted for our proposed with the

Fig. 5 Single user community in small-scale network. **a** Single-user community with redundant users. **b** Single-user community with reduced redundant users

Fig. 6 Removal of redundant user profiles in LSN. **a** Multi-user community with redundant users. **b** Multi-user community with reduced redundant users

Fig. 7 Analysis of remaining number of users after merging redundant user identities in SSN & LSN

standard Profile Matching Technique [4] and state-of-the-art model, namely ID-Connect [12]. The results in Fig. 7c show the comparatively better capability of our proposed identity resolution model than both the methods.

Our experimentation is further extended to a hashtag-based Massive Twitter Network (MTN) generated from a Twitter trend. In order to construct the trend-based network, the Twitter cursor is used to perform the REST Application Programming Interface (API) search for tweeted posts that contained the trend #FakeNews. The tweets are fetched in JavaScript Object Notation (JSON) format which consists of 18,000 posts. This is followed by deriving the list of distinct users who participated in the dialogue in the form of tweets or retweets. Further,

the follower-following relationship existing among these distinct users is extracted and trend-based MTN is plotted as highlighted in Fig. 8. The directed Twitter graph constructed from the real-world trend contains a total of 23,000 nodes and 36,118 edges. However, owing to its dense-overlapped arrangement of nodes, the detected 150 users having redundant profiles are graphically illustrated separately in Fig. 9, instead of highlighting them in MTN. The results in Fig. 9 additionally show a comparative performance of our proposed approach with the other existing methods. After the fifth iteration, our proposed model and ID-Connect shows stability, while the profile matching method is found to be unstable. This means that either profile matching technique requires more iterations

Fig. 8 Trend-based massive twitter network (MTN)

to attain stability or might remain unstable by incorrectly categorizing authentic user profiles as redundant. The results endorse better performance of our proposed identity resolution model as the network size scales over an increasing number of nodes and edges. Another interesting result highlighted in Fig. 10 plots the computed accuracy for MTN. This clarifies that though ID-Connect converges in a lesser number of iterations as similar to our

method, but gives lesser accuracy than our proposed identity resolution model.

6 Conclusion

A person may have multiple accounts on online social media networks depending upon specific usages, such as for professional connection they use LinkedIn, for group discussions, one may prefer Facebook, Google+, etc. However, some malicious users, mask their original identity and tend to perform deceiving attempts, either to steal private information or cyberbully online users. This necessitates resolving the multiple presences of malicious users over social media. Identity resolution remains an ever-challenging task that involves finding redundant users by matching profiles from multiple social network profiles. Identity resolution conducted based on real-world data extracted from social media can be useful in several fields such as recommendation, searching, and integration of profile, false event, information diffusion, etc.

Work discussions In our research, we have focussed on the identity resolution problem in the online social network, where a user tends to create a profile with multiple attributes, like username, screen name, user description, location, etc. These multiple attributes form the profile identity of the specific user, which could be further used for matching similarity among other existing profiles. Profile details are found by our proposed identity search algorithm. We have also designed an exploratory algorithm to perform profile matching of users by applying

Fig. 9 Comparative performance analysis all models

Fig. 10 Comparison of model accuracy

similarity computation on attribute values. This is followed by executing identity linking and merging algorithm for removing the redundant impact of the repeated profiles. To show the effectiveness of our algorithms, we have collected user profile details to form small-scale as well as a large-scale user-centric Twitter social network. Our proposed identity resolution algorithm was executed on the user-centric network for identifying users with redundant profiles. The experimentation results showed the successful merging of similar profile. By this process, the existing social network reduced in several nodes, as redundant user-profiles merged, thereby having only authenticated users. Moreover, our model is applied to a massive Twitter network with a comparative study showing better results as compared to its existing counterparts.

Limitation A lot of effort is required to fetch data from the Twitter online social network via public API. Recently social media has applied various restrictions on data retrieval. We can retrieve data of online users up to a certain limit. Due to the own data privacy of online social networks, the private data of users is not accessible. Therefore it is imperative to develop a multi attribute identity resolution method to a large extent efficiently.

Future research directions This work can be extended to a single user across multiple platforms and multiple users across multiple platforms. Improving the similarity techniques could be explored for even better results to reveal and eradicate malicious user-profiles and other details. Exploring diffusion more efficiently on the reduced

graph, to study the pattern of information, could open yet another avenue for future research.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

- Jain P (2015) Automated methods for identity resolution across heterogeneous social platforms. In: HT'15 Proceedings of the 26th ACM conference on hypertext and social media, pp 307–310
- Jain P, Kumaraguru P, Joshi A (2013) @I seek 'fb.me': identifying users across multiple online social networks. In: International world wide web conference committee (IW3C2), pp 1259–1267
- Soltani R, Abhari A (2013) Identity matching in social media platforms. In: International symposium on performance evaluation of computer and telecommunication systems (SPECTS), pp 64–70
- Raad E, Chbeir R, Dipanda A (2011) User profiles matching in social networks. In: NBIS'10 Proceedings of the 2010 13th international conference on network-based information systems, pp 297–304
- Malhotra A, Totti L, Meira W Jr, Kumaraguru P, Almeida V (2013) Studying user footprints in different social networks. In: ASONAM'12 Proceedings of the 2012 international conference on advances in social networks analysis and mining (ASONAM 2012), pp 1065–1070

6. Mugan J, Chari R, Hitt L, McDermid E, Sowell M, Qu Y, Coffman T (2014) Entity resolution using inferred relationships and behaviour. In: International conference on big data, pp 555–560
7. Wang Z, Zhou C, Sun J, Wang S, Zhan H, Yu Y, Zhu W, Cui X (2015) Key technology research on user identity resolution across multi social media. In: International conference on cloud computing and big data, pp 358–361
8. Keretna S, Hossnv A, Creighton D (2013) Recognizing user identity in twitter social network via text mining. In: IEEE international conference on systems, man and cybernetics, pp 3079–3082
9. Edwards M, Wattam S, Rayson P, Rashid A (2016) Sampling a labelled profiled data for identity resolution. In: IEEE international conference on big data, pp 540–547
10. Lai S, Li H, Zhu H, Ruan N (2015) De-anonymizing social networks: using user interest as side-channel. In: IEEE/CIC international conference on communication in China (ICCC)
11. Torky M, Meligy A, Ibrahim H (2016) Recognizing fake identities in online social networks based on a finite automaton approach. In: 12th international on computer engineering conference (ICENCO), pp 1–7
12. Azad MA, Morla R (2015) ID-CONNECT: combining network and call features to link different identities of user. In: IEEE 18th international conference on computational science and engineering (ICCSE), pp 160–167
13. He D, Cao Z, Dong X, Shen J (2014) User self-controllable profile matching for privacy-preserving mobile social networks. In: IEEE international conference on communication system (ICCS), pp 248–252
14. Iofciu T, Fankhauser P, Abel F, Bischoff K (2011) Identifying users across social tagging systems. In: Proceedings of the fifth international AAAI conference on weblogs and social media, pp 522–525
15. Bartunov S, Korshunov A, Park ST, Ryu W, Lee H (2012) Joint link-attribute user identity resolution in online social network. In: 6th SNA-KDD workshop'12
16. Reihanian A, Minaei-Bidgoli B, Alizadeh H (2016) Topic-oriented community detection of rating-based social networks. *J King Saud Univ Comput Inf Sci* 28(3):303–310
17. Ali M, Yaacob RA, Endut MN, Langove NU (2017) Strengthening the academic usage of social media: an exploratory study. *J King Saud Univ Comput Inf Sci* 29(4):553–561
18. Wani MA, Jabin S (2018) Mutual clustering coefficient-based suspicious-link detection approach for online social networks. *J King Saud Univ Comput Inf Sci*. <https://doi.org/10.1016/j.jksuci.2018.10.014>
19. Somani S, Jain S (2017) Resolving identities on Facebook and Twitter. In: 10th IEEE international conference on contemporary computing (IC3), pp 1–3
20. Ahmad W, Ali R (2019) Social account matching in online social media using cross-linked posts. *Procedia Comput Sci* 152:222–229
21. Li J, Wang AG (2015) A framework of identity resolution: evaluating identity attributes and matching algorithms. *Secur Inform* 4:6
22. Srivastava DK, Roychoudhury B, Samalia HV (2018) Importance of user's profile attributes in identity matching across multiple online social networking sites. In: 8th IEEE international conference on cloud computing, data science and engineering (confluence)
23. Sharma V, Dyreson C (2018) LINKSOCIAL: linking user profiles across multiple social media platforms. In: IEEE international conference on big knowledge (ICBK)
24. Kumar P, Sinha A (2016) Real-time analysis and visualization of online social media dynamics. In: IEEE international conference on next generation computing technologies, Dehradun, pp 362–367
25. Zhang H, Kan MY, Liu Y, Ma S (2014) Online social network profile linkage. In: Asia information retrieval symposium. Springer, Cham, pp 197–208
26. Jamjuntra L, Chartsuwan P, Wonglimsamut P, Porkaew K, Supasitthimethee U (2017) Social network user identification. In: 9th IEEE international conference on knowledge and smart technology (KST), pp 132–137
27. Alvarez JJ, Mendoza FA, Labrador M (2017) An accurate way to cross reference users across social networks. In: IEEE conference SoutheastCon
28. Malviya V, Gupta GP (2015) Performance evaluation of similarity-based link prediction schemes for social networks. In: IEEE international conference on next generation computing technologies, Dehradun, pp 654–659

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.