Identifying Users from Online Interactions in Twitter

Madeena Sultana(✉), Padma Polash Paul, and Marina Gavrilova

Department of Computer Science, University of Calgary,
2500 University DR NW, Calgary, Canada
{msdeena, pppaul, mgavrilo}@ucalgary.ca

Abstract. In recent years, the mass growth of online social networks has introduced a completely new platform of analyzing human behavior. Human interactions via online social networks leave big trails of behavioral footprints, which have been investigated by many researchers for the purpose of targeted advertising and business. However, analysis of such online interactions is rarely seen for user identification. The main objective of this paper is to analyze individuals' online interactions as biometric information. In this paper, we investigated how online interactions retain behavioral characteristics of users and how consistent they are over time. For this purpose, we proposed a novel method of identifying users from online interactions in Twitter. Identification performance has been evaluated on a database of 50 Twitter users over five different time periods. We obtained very promising results from experimentation, which demonstrate the potential of online interactions in aiding the authentication process of social network users'.

Keywords: Social behavioral biometrics · Online social networks · Cyber security · Person authentication · Biometric recognition · Knowledge discovery · User profiling

1 Introduction

User identification is the key requirement for secured access of information and service. The traditional user authentication mechanisms are based on PIN, Passwords, or identity cards. However, reports on security of traditional identification systems point out how easy it is nowadays to break majority of "strong" passwords or PINs [5–7]. Moreover, passwords, PINs, or identity cards can easily be shared, stolen, or forgotten. Biometric-based user identification systems can overcome such drawbacks by using physiological or behavioral characteristics of individuals. Physiological characteristics such as face, fingerprints, iris etc. and behavioral characteristics such as voice, gait, signature etc. cannot be shared, stolen or forgotten as well as quite difficult to change or forge. Although behavioral biometrics are more volatile to changes one undergoes through the life time, they offer advantages of being dynamic, non-intrusive and cost effective over physiological biometrics [3, 4]. Behavioral biometrics are difficult to fake or imitate because of its dynamic nature. For instance, impersonating a person's walking style or gait is way more difficult than creating a fake fingerprint. For such advantages,

M.L. Gavrilova et al. (Eds.): Trans. on Comput. Sci. XXVI, LNCS 9550, pp. 111–124, 2016.
DOI: 10.1007/978-3-662-49247-5_7

behavioral biometrics are becoming a popular alternative to well-established physiological biometrics in many authentication applications to reduce security threats, especially in the cyberworld [1–3]. In [8], we introduced a novel biometric trait called Social Behavioral Biometrics (SBB), which is based on the hypothesis that a person can be identified from social activities and interactions. At present, with the advent of social online social networks, human behavior has been expanded in virtual world. In [9], we identified online social media as the source of social behavioral biometrics. This paper validates the idea by identifying behavioral patterns in online interaction of users, evaluate their performance for user identification. In other words, our aim is to investigate the underlying patterns of user interactions in social networking platforms and evaluate their discriminability for being used in identification purpose.

The increasing popularity of social networking websites such as Facebook, Twitter, Myspace, LinkedIn, Flickr, etc. has turned them to massive sources of 'big data', which offer new opportunities of studying human behaviors from different perspectives. In fact, they are considered as one of the most valuable, diversified, and dynamic repository of information which can avail data from health sector to targeted advertising [10]. Human behavior and social relationships via social networking websites have been studied by many researchers for many different applications such as targeted marketing, recommendation systems, prediction of stock market, and so on. Twitter is one of the most studied social networking website from diverse fields of research. Twitter is a platform, which enables users to interact with other users through connections. According to a statistical report [11], Twitter has around 5.5 billion of active registered users who produces 58 million tweets per day and 9,100 tweets per second. In addition, 135,000 new users are signing up to twitter every day. This statistics demonstrates that billions of users nowadays are communicating to each other through Twitter. In Twitter, users communicate by posting real time micro-blogs called Tweets. Unlike other mediums, tweets are restricted to 140 characters of length. Despite limitation of size, tweets are rich sources of information about the user and his communicative behavior. Tweets not only comprise of texts but also contain interactions to other users or websites such as replies, retweets, URLs etc. Therefore, we categorize tweets into two types: textual and interactive. Textual parts are sometimes called original tweets [8] that are produced by the users targeting the general audience. Interactive tweets possess some kind of communicative materials such as replies, retweets, URLs, and hashtags. Replies are tweets intended to another specific user. Replies are directed to a specific user by appending @ symbol at the beginning of the intended user name (e.g. @user_abc). The main difference between reply and original tweets is - replies are intended to a specific user whereas original tweets are more like broadcasting messages to all instead of addressing a particular user. Replies could be very different from original tweets since a person may alter their writing style when addressing another person [12]. Retweets are tweets, originally written by another user but posted or shared by the user in his/her timeline. It is an act of sharing another user's tweets to all of the followers of a user [13]. Another distinctive feature of tweets is the hashtags. Hashtags are shorthand convention adopted by Twitter users for assigning their posts to a wider corpus of messages on the same topic [14]. We have considered hashtags as an interactive feature of a user to a wider corpus. Later in this paper, we would show how personal topics of interest could be explored by analyzing hashtags.

The last interactive features we identified in tweets is shared URLs. Users can direct his followers to a news article, a blog post, related websites, videos, or photographs by sharing the URLs. In this paper, we will investigate whether it is possible to identify some behavioral patterns from such interactive data of Twitter users. Therefore, the main contribution of this paper is to identify some behavioral features from interactive Twitter data of users and formulating a framework for user identification by matching such behavioral patterns. We would extend our study to analyze the impact of feature matching in different time intervals in order to evaluate how consistent such behavioral characteristics are.

2 Literature Review

The security in cyberworld is important as in the real world [4, 15]. In some cases, it is more crucial since breaching the security in cyberspace may jeopardize our life beyond monetary loss. Therefore, person authentication plays an important role in fortifying the security of virtual world. Behavioural biometrics are more popular for person authentication in virtual world rather than in the real world. Many new behavioral biometrics have been proposed within the last few years. Some state-of-the-art works are described below to analyze the current trend of the behavioral biometric research.

In 2011, mouse dynamics have been studied as behavioral biometrics by Jorgensen and Yu [16]. Instead of ordinary static authentication, keystroke dynamics have been utilized for continuous authentication by Bours [17] in 2012. In this work, real time typing pattern of the user is continuously matched with the stored template, which is referred as continuous authentication. Once the matching score or trust level decreases below a certain threshold, there is a possibility of the system of being locked. Frank et al. [18] identified 30 behavioral touch features from raw touchscreen logs of smartphones and named these novel biometric features as Touchalytics. This study demonstrates that different users have distinct pattern of navigation and this behavioral patterns exhibit consistency over time. Guo et al. [19] explored that person can be identified by his/her own style of 'handshaking'. Handshaking is a specific set of human actions, needed to unlock the screen of the smartphone of a user. The authors of [19] observed unique, stable, and distinguishable idiosyncratic patterns in handshaking behaviors of users and used them to authenticate users of their smartphones. In [20], Feng et al. identified biometric characteristics in mobile device picking-up motion. A novel behavioral biometric called hobby driven biometric has been introduced by Jiang et al. [21]. In this work, a comprehensive study on habitual behaviors driven by hobbies is presented. Considering the decorating and tidying style of a room as hobby-driven behavior, the authors conducted a survey on 225 people of different ages and professions. They observed unique and steady characteristics based on style, color, position, and habitual operating order of the object for different persons. This study demonstrates that a person's own choice and habits possess enough discriminability to act as behavioral biometric features. Another study on person's web browsing style has been conducted by Olejnik and Castelluccia [22]. The authors investigated web browsing data of 4,578 users and revealed idiosyncratic patterns of browsing style. The authors claimed that person authentication, anomaly, and fraud detections etc.

would be the potential applications of their browsing style-based behavioral biometric trait. A multimodal behavioral biometric system by combining inputs from keyboard, mouse, writing sample, and web browsing history at decision level has been proposed by Fridman et al. [23]. Another state-of-the-art work by Bailey et al. [24] presents a multimodal behavioral system combining data from keyboard, mouse, and Graphical User Interface (GUI) interactions. Bailey et al. [24] experimented fusion at feature and decision level, where decision level fusion obtained higher recognition performance than feature level fusion.

From the above summary, it is pertinent that idiosyncratic patterns can be found in every aspect of human actions ranging from walking style in real world to browsing style in cyberworld. Therefore, human activities on online social networking platforms should have some patterns for being used as biometric features for person authentication. This has motivated us to study the users' social interactions in Twitter to explore social behavioral biometric features for person identification.

3 Proposed Method

In this paper, we are interested to reveal behavioral characteristics from user interactions via Twitter. Our assumption here is each individual has his own pattern of interactions in social networks such as he has his own set of friends whom he contact regularly, some topics of interests, some preferred websites and so on. Therefore, mining interaction-based data of individuals may reveal the very personal characteristics of a user, which can eventually be used as personal signature. In this paper, we proposed a framework to identify Twitter users based on their dynamic communication behavior rather than analyzing of their static information. Our proposed methodology has the following steps:

(i) Acquire tweets from selected user profiles.
(ii) Extract interactive data such as replies, retweets, hashtags, and URLs from tweets.
(iii) Analyze acquired social data to explore social behavioral biometric features.
(iv) Apply matching techniques on training and test set for person authentication.

Figure 1 shows a basic block diagram of the proposed method. The following subsections contain detailed description of the aforementioned steps.

3.1 Data Acquisition

We collected tweets of 50 users over a period of more than four months. Initially, 300 active users from Twitter are selected on a random basis who uses English for writing micro-blogs or tweets. Among them, the most prolific 50 users are identified who produces more than 100 tweets per week on average. We have divided our data collection period into five sessions separated by at least 20 days interval to avoid possible overlapping in data. In this paper, we denote session as S_i where $i = 1, 2, 3, 4, 5$. In every session, approximately 200 tweets of each user are collected and stored.

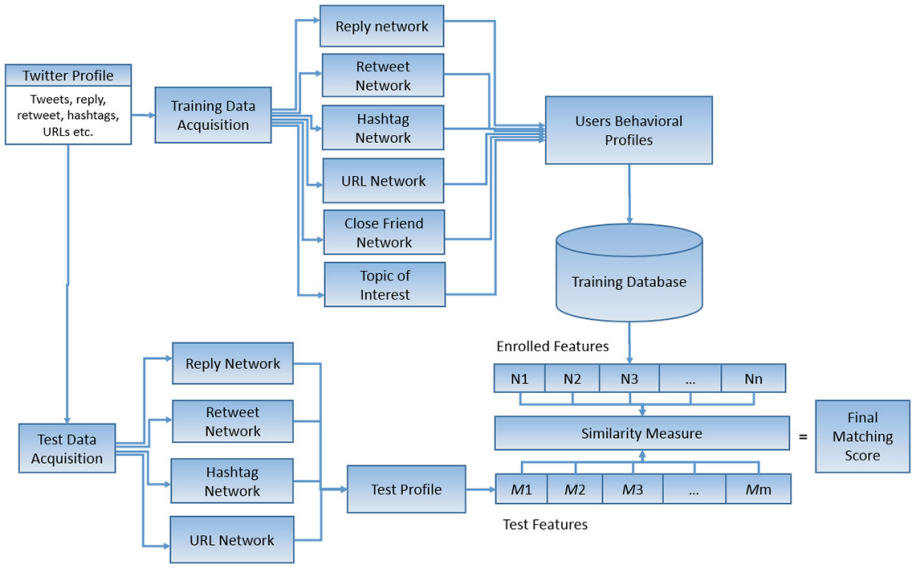

Fig. 1. Block diagram of the proposed framework of Twitter user identification using online behavioral profile.

However, the collected data is unbalanced due to limitation of Twitter API. Our collected dataset includes retweets, replies, hashtags, and URLs along with texts.

Since our goal is to identify behavioral pattern in dynamic communication of users, we considered the interactive or communicative data such as replies, retweets, URLs, and hashtags. In tweets, usually replies are preceded by @, retweets are preceded by RT, and hashtags are preceded by # symbol. Individuals also post URL(s) to share more information with other users that they could not accommodated in timelines. The key challenge for framing person authentication from social interactive data is the selection of feature set that best describes the behavioral pattern of users. For this purpose, retweets, replies, hashtags, and URLs are extracted from the dataset of each user from every session. We also extracted some statistics such as number of tweets, retweets, retweeted persons, replies, replied persons, URLs, distinct URLs, hashtags, and distinct hashtags per week. Unlike other biometric traits, the behavioral characteristics are not readily available in social data. Therefore, social data is needed to be analyzed to extract behavioral features. The feature extraction technique is described in the following subsection.

3.2 Feature Extraction

The communication-based features from Twitter data are analyzed in this paper. Social data of each session is analyzed to discover distinguishable personal characteristics of users. Analysis has been done in two steps.

Initially, extracted social data from each session is analyzed based on frequency. The reason for frequency analysis is that frequent interactions of a user indicates consistent behavioral pattern. For instance, if user A frequently retweets user B, then user A has a strong preference of retweeting user B. Therefore, frequent social communicative data are extracted from replies, retweets, hashtags, and posted URLs of each session. Networks of communication are formed based on such frequent data. Frequency analysis is accomplished on the social data from the first four sessions (S_1, S_2, S_3, and S_4). Four networks are created from each of the data set of S_1, S_2, S_3, and S_4 to represent frequency-based social behavioral biometric features. These four networks are reply network, retweet network, hashtag network, and URL network. In this paper, reply network, retweet network, hashtag network, and URL network of session i is denoted as R_{Si}, RT_{Si}, H_{Si}, and L_{Si}, respectively.

Biometric feature extraction from Twitter data is an evolving process. Knowledge can be discovered by analyzing data over time. Therefore, our next step is to investigate the interaction-based Twitter data over time to explore consistent behavioral patterns. In this paper, we are proposing two knowledge-based behavioral features, which can be extracted by analyzing the frequency-based networks and interactive social data. Our proposed knowledge-based features are close friend network and the topics of interest. Figure 2 shows the hierarchy of frequency-based and knowledge-based features created from social data of four sessions.

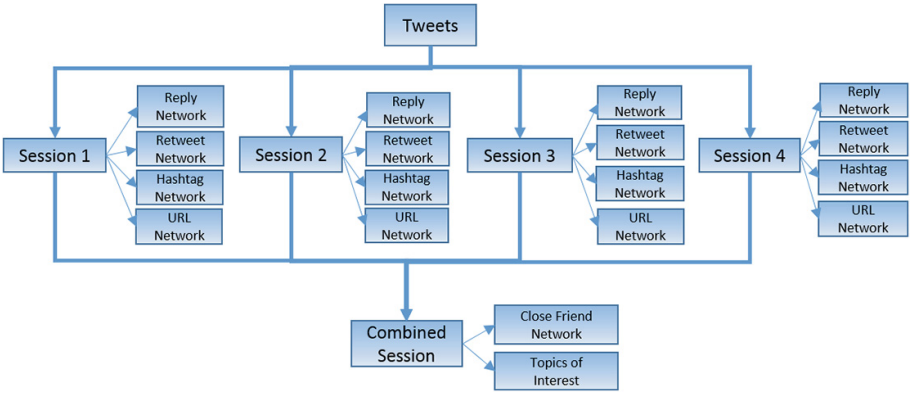

Fig. 2. Hierarchy of frequency-based and knowledge-based features from first four sessions.

The proposed frequency-based and knowledge-based behavioral features are explained hereafter.

- **Reply Network:** Reply network is generated by analyzing the replies of a specific user in Twitter. In this network, nodes are the user him/herself and the other users he/she replies. The process includes three major steps: listing all persons that the user replied to, counting the replies to each person, and finally applying a threshold. A threshold is applied on the counts of the replies for each individual the user replied to, since we are interested to include persons in the network whom the user

replies frequently. A person is added to the reply network if and only if the count of reply is greater than a certain threshold. This network explores a short list of friends/persons whom the user communicates the most frequently.

- **Retweet Network:** A retweet network is created similarly to the reply network to find the most frequent retweeted persons in the user's timeline. However, person's retweeting behavior is more likely to be changed over time. For our study, we categorized retweets based on the direction: retweeting and retweeted. Retweeting refers that a person is posting tweets of other users in his/her timeline. On the other hand, retweeted means the person's tweet is retweeted by some other users or acquaintances. Both are related to the proposed retweet network with the differences of direction only. Based on the direction of the retweets two separate networks are constructed. We investigated both ways retweets to explore some consistency of the users retweeting behavior. For this purpose, we listed all the persons who are present in both retweeted and retweeting network for each session. The idea is, if user A retweets B then we also investigated whether B retweets A. This relation strengthens the retweeting relationship among A and B that is less likely of being random. Finally, retweet network is constructed using frequent items of retweeting network and common items of retweeted and retweeting networks.
- **Hashtag Network:** A hashtag network is created similarly to reply network. In this network, all hashtags that qualify over a certain threshold are added as nodes in the network. However, hashtag network should never be used as a static feature. It should be updated after a certain amount of time to cope with the changes of interest of the person over time.
- **URL Network:** URLs are often shared by users in micro-blogs. Building a URL network may reveal personal interests and URL sharing pattern of users. The creation of URL network is similar to reply network.
- **Close Friend Network:** Reply and retweeting activities of users are likely to be changed over time due to various events of life. For example, job switching of a person may initiate active communication to a group of new friends (i.e. colleagues of new job) and communication to some of the early acquaintances (i.e. colleagues of previous job) may disappear. However, some of the acquaintances consistently remain in user's timeline over longer period or may appear after some irregular intervals. We have considered such acquaintances as the close friends of individuals in social media. Creating a network of such close friends and later using it as a social biometric feature would enhance the reliability of the authentication process. Therefore, we analyzed retweet and reply networks to identify close friends of each of our 50 subjects. At first, we listed all the persons who are present in either retweet or reply network for at least two sessions. Here the fact is that if an acquaintance appears regularly in the user's timeline in form of reply or retweets, that acquaintance should have strong relationship with the user. Further strong communicative relationships are explored by combining people who exist in both reply and retweeted network of each session. Finally, a close friend network is created by using all the above listed distinct persons as nodes. It is worth mentioning that these networks would be more consistent if the aforementioned analysis could be conducted over a longer period with more sessions.

- **Topics of Interest:** Hashtag sharing of a Twitter user is an event driven activity and the most likely feature to be changed over time. However, analyzing hashtags over longer period may reveal consistent information about the user such as personal interest. For example, if some hashtags appear regularly in the tweets of a user over long period, then the user is consistently interested about that issue. Such pattern is an important feature to identify the user in virtual domain. The problem of topic analysis in hashtags is that many different hashtags might represent a single topic. Therefore, relevant hashtags have to be clustered to explore a more general topic of interest. Some examples of similar hashtags and their general topic of interest are presented in Table 1. We accumulated all hashtags shared by each user from four sessions. Then, all hashtags are clustered according to the similarity of words. For example, all hashtags presented in the second row of the Table 1 contain a common word "climet". Therefore, we assigned "climet" as one of the topics of interest for that person. Any hashtag shared by that particular user containing "climet" would fall under this topic of interest. Any cluster containing a single element is removed from the list. Finally, we explored consistent topics of interest from the list by discarding topics present in only a one session since such topics are more likely to be random and event driven. In this way, we assigned some general topics of interest for each person by clustering similar hashtags. It is also worth mentioning that clustering is accomplished based on string similarity not semantic similarity of hashtags. Further analysis could be done by finding any correlation between hashtags and shared URLs or replied users.

Table 1. Example of assigning topic of interest for similar hashtags

Similar hashtags	Topic of interest
#Climet, #ClimetLeaders, #ClimetReality, #ClimetChange	climet
#Immigration, #Immigrationreform, #Immigration&hellip, #iamimmigration &amp	immigration
#Sydneyhalfmarathon, #Sydney, #Sydneyweather, #Sydneywinter, #SydneyFires	sydney

3.3 Similarity Matching

For each user in our data set, we created reply network, retweeted network, retweeting network, hashtag network, URL network for each of the first four sessions. Then the close friend network and topics of interest are extracted by combining data from more than one sessions. Feature sets from the first four sessions are considered as training sets. The fifth session is considered as test data set. Reply, retweeted, retweeting, hashtag, and URL networks are generated from the fifth session as well to form features of test set. The difference between test and training networks is that no threshold values are applied during the test networks generation. Initially, matching is accomplished on corresponding reply, retweeted, retweeting, URL, and hashtag networks of test and training sets. Then, test hashtag network is matched with the topic of interest of training

sample. Also, close friend network of the training sample is matched with the combined reply and retweeted networks of the test sample. The features of training and testing networks are matched to obtain the final similarity score of a person's identity. The matching of hashtag network of the test set and topics of interests of the training set is accomplished using Levenshtein distance [25]. The similarity scores of the other five test and training networks are calculated as the ratio of the number of mutual nodes in test and training networks to the number of nodes in training network. Therefore, the similarity score, S_p is defined as follows:

$$S_p = \frac{|P_T \cap P_R|}{|P_T|} where\, P_T \neq \emptyset \tag{1}$$

Where training and test networks are represented by P_T and P_R respectively. The similarity scores are then normalized and fused to obtain the final matching score. The framework of our feature extraction and matching system is implemented in such a way that we can feed more features in future. The performance of the proposed feature sets for person identification is demonstrated in the following section.

4 Experimental Results and Discussion

During experimentation, we aimed to evaluate the performance of the extracted features for person identification. Performance variations of the proposed features over time were another point of interest of our experiments. Therefore, we performed three sets of experiments to evaluate the performance proposed method. In the first set of experiments, we evaluated the performances of each features and their combined forms. In the second set of experiments, we compared performance of all frequency-based features in different sessions. Finally, in the third set of experiments, we compared performance of all features considering more than one training sessions. All experiments were carried out on Windows 7 operating system, 2.7 GHz Quad-Core Intel Core i7 processor with 16 GB RAM. Matlab version R2013a was used for implementation and experimentation of the proposed method. We collected social data of 50 users from five different sessions. The first four sessions were used as training sets and the fifth session was considered and test set. Table 2 shows training and test set separation of collected Twitter data in different sessions.

Table 2. Training and test sets of collected Twitter data in different sessions.

Session	Number of users	Session interval	Training/test
Session 01	50	0 days	Training
Session 02	50	20 days	Training
Session 03	50	20 days	Training
Session 04	50	20 days	Training
Session 05	50	20 days	Test

In the first experiment, we evaluated the performance of each single feature and their combined forms. As discussed in the feature extraction section, we have two types of feature sets: frequency-based and knowledge-based. In this experiment, a particular feature from the test set of each subject is matched with the corresponding feature of all subjects in the training set. Figure 3 plots the performance of the proposed features in terms of Receiver Operating Characteristics (ROC) curves. ROC curve plots False Rejection Rate (FRR) and False Acceptance Rate (FAR) with respect to different threshold values. Equal Error Rate (EER), the optimal point on ROC where FAR is equal to FRR, is often considered as a measure of identification or verification performance [26]. The training set consists of reply, retweet, URL, hashtag, close friend network, and topics of interest. From Fig. 3, one can see that EER is high for single feature. However, knowledge-based features i.e. close friend network and topics of interest have better performance than frequency-based features i.e. reply, retweet, hashtag, and URL networks. The average EER of the four frequency-based features is around 35 %. On the other hand, the average EER of the two knowledge-based features is around 28 %. This finding confirms the consistency of the knowledge-based features over time. Figure 3 also shows that the EER reduced significantly while all frequency-based features are combined. The best identification performance is achieved by combining all six features. In this case, the EER is as low as 20 % approximately. However, the choice of the optimal point may be altered according to security level [26]. For example, a point where the FAR is low and the FRR is high is suitable for high security applications. Alternatively, a point with low FRR and high FAR is good for low security applications where the system may allow a reasonable amount of false alarms.

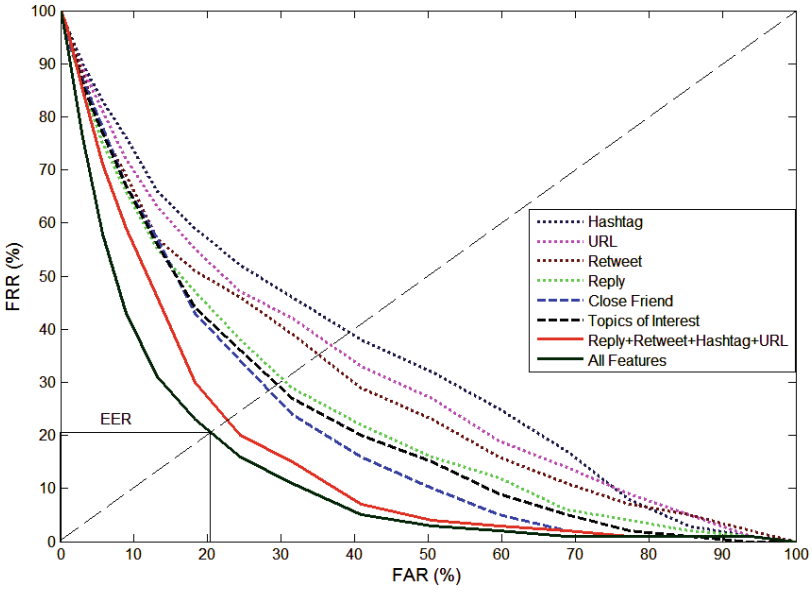

Fig. 3. ROC curves of person identification using proposed features.

As per the discussion in the feature extraction section, frequency-based feature sets, collected within a short span of time, might be subjects of changes over time. Therefore, we are also interested in investigating the effects of changes of the proposed frequency-based features for person identification. We set up the second experiment to evaluate the performance of combined frequency-based features from different sessions. In this case, the test set is consists of reply, retweet, hashtag, and URL networks of 50 users from session 5. The test feature set is then matched with the reply, retweet, hashtag, and URL networks from session 1, 2, 3, and 4, sequentially. The sessions are numbered according to the sequence of data acquisition. Therefore, session 5 contains the most recent social data of the users whereas session 1 comprises of the least recent data. Figure 4 plots the identification results of the four sessions in terms of ROC curves. It has been observed from Fig. 4 that the best identification result (EER = 24 %) is obtained by measuring the similarity between the closest test and training session pairs i.e. session 5 and session 4. A little degradation in performance has been observed when the test data is matched with session 2 and session 3. However, the EER has been increased to around 33 % for the similarity measure between session 5 and session 1. In this case, the training data set (session 1) is more than four months older compared to the test set (session 5). From the second experiment, one can see that frequency-based features need to be updated regularly to cope with the changing social behavior of a person.

In the third set of experiments, we evaluated the performance of the proposed method using more than one training samples. The experiment was conducted using the

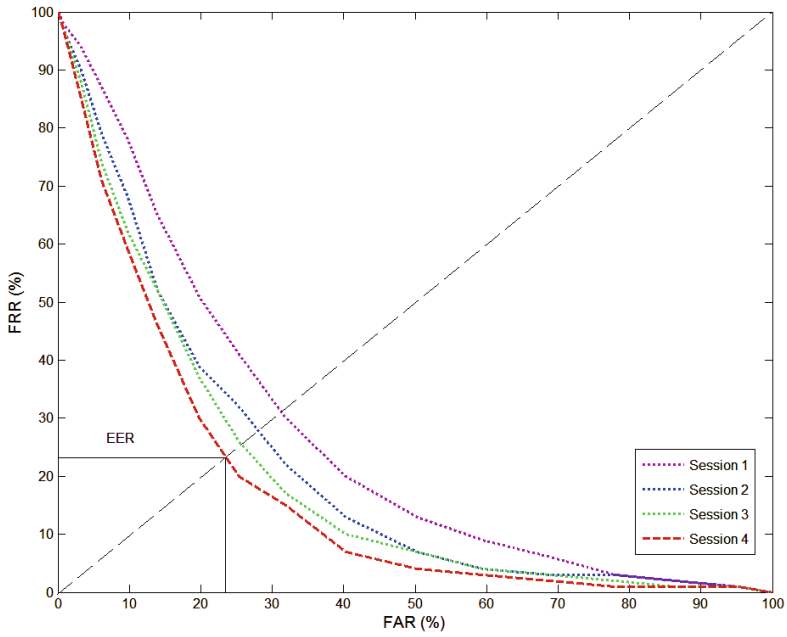

Fig. 4. ROC curves of combined frequency-based features in different sessions.

first two, three, and four sessions as training sets in an incremental order. Therefore, in this experimentation each feature has more than one samples, unlike the first two set of experiments. However, session five is used as the test set, similar to the aforementioned experiments. Figure 5 plots the ROC curves of the combined sessions. From Fig. 5, one can see that consideration of more than one training samples enhanced the performance of the proposed method. Compared to single training session as shown in Figs. 3 and 4, use of combined training features sets as shown in Fig. 5 obtained less EER. The lowest EER = 18 % has been obtained using four training sessions. This experiment demonstrate that all the sessions contain some consistent feature sets. This is why, having more than one sessions in training set reinforces consistent features and increases recognition performance.

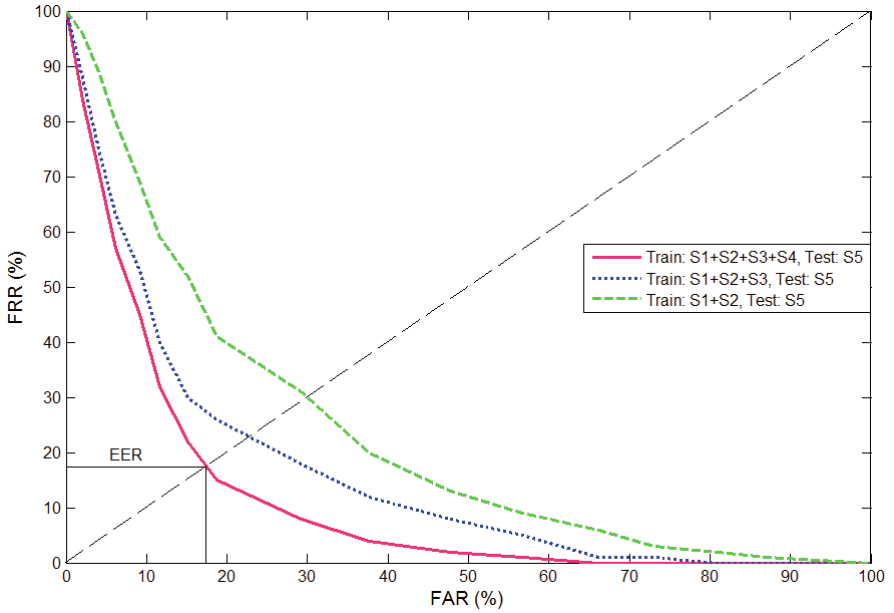

Fig. 5. ROC curves of the proposed method using 2, 3, and 4 training sessions in an incremental order.

Our experimental results demonstrate that the online social interactions of a person such as replies, retweets, URLs, hashtags, close friends, topics of interest possess idiosyncratic patterns, which can be used as biometric signature of that person. It has also been demonstrated that consistent behavioral patterns can be explored by applying knowledge-based analysis and data mining techniques on social data. Such knowledge-based features can enhance the performance of a biometric person authentication system. Similar to other behavioral biometrics, the proposed features are also subjects to changes over time. Therefore, they are needed to be updated periodically. The recognition performance might be degraded otherwise. However, this requirement should not pose any problem, since social behavioral biometrics can easily

be extracted and updated periodically without interfering the users at all. One limitation of our experimentation is that the investigation has been conducted on the data of relatively small number of subjects. We believe that more data would help to further support our findings.

5 Conclusions

In this paper, social interactions via Twitter has been analyzed to discover behavioral signature of users. We introduced frequency-based and knowledge-based behavioral biometric features, which can be extracted from any online social networking platforms. Our experimentation includes performance and consistency evaluation of the proposed behavioral biometric features for closed set person identification. There are three important findings from this research. Firstly, online communication or interactions of users retain behavioral footprints of individuals. Secondly, analyzing social data over a period of time can explore underlying behavioral pattern, which exhibits strong idiosyncratic characteristics. Finally, such behavioral patterns of users maintain stability over time to some extent and do not change overnight. Experimental results demonstrate encouraging performance of using online interaction-based features for user identification. The applications of the proposed social behavioral biometrics features could be as diverse as person authentication, access control, anomaly detection, customer profiling, behavior analysis, situation awareness, risk analysis, friend recommendation systems, and so on. Our future research includes expanding the concept to include broad range of on-line social communications and environments.

Acknowledgement. The authors would like to thank NSERC DISCOVERY program grant RT731064, URGC, NSERC ENGAGE, NSERC Vanier CGS, and Alberta Ingenuity for partial support of this project.

References

1. Moskovitch, R., Feher, C., Messerman, A., Kirschnick, N., Mustafic, T., Camtepe, A., Lohlein, B., Heister, U., Möller, S., Rokach, L., Elovici, Y.: Identity theft, computers and behavioral biometrics. In: Proceedings of IEEE International Conference on in Intelligence and Security Informatics (ISI 2009), pp. 155–160. IEEE (2009)
2. Gavrilova, M.L., Monwar, M.: Multimodal biometrics and intelligent image processing for security systems. IGI Global (2013)
3. Yampolskiy, R.V., Govindaraju, V.: Behavioral biometrics: a survey and classification. Int. J. Biometrics **1**(1), 81–113 (2008)
4. Yampolskiy, R.V., Gavrilova, M.L.: Artimetrics: biometrics for artificial entities. IEEE Robot. Autom. Mag. **19**(4), 48–58 (2012)
5. Monwar, M., Gavrilova, M.L.: Multimodal biometric system using rank-level fusion approach. IEEE Trans. Syst. Man Cybern. B Cybern. **39**(4), 867–878 (2009)
6. Zhang, H., Li, M.: Security vulnerabilities of a remote password authentication scheme with smart card. In: Consumer Electronics, Communications, and Networks, pp. 698–701 (2011)

7. Paul, P.P., Gavrilova, M., Klimenko, S.: Situation awareness of cancelable biometric system. Vis. Comput. **30**, 1–9 (2013)
8. Sultana, M., Paul, P.P., Gavrilova, M.: A concept of social behavioral biometrics: motivation, current developments, and future trends. In: CW Biometric Workshop (2013)
9. Sultana, M., Paul, P.P., Gavrilova, M.: On-line user interaction traits in web-based social biometrics. IGI Chapter, pp. 177–190 (2014)
10. Bringmann, B., Berlingerio, M., Bonchi, F., Gionis, A.: Learning and predicting the evolution of social networks. Intell. Syst. **25**(4), 26–35 (2010)
11. http://www.statisticbrain.com/twitter-statistics/. Accessed 25 March 2015
12. Grant, T., Laboreiro, G., Maia, B., Oliveira, E., Sousa Silva, R., Sarmento, L.: 'twazn me!!!; (' automatic authorship analysis of micro-blogging messages. In: Muñoz, R., Montoyo, A., Métais, E. (eds.) NLDB 2011. LNCS, vol. 6716, pp. 161–168. Springer, Heidelberg (2011)
13. https://support.twitter.com/articles/166337-the-twitter-glossary. Accessed 25 March 2015
14. Carter, S., Tsagkias, M., Weerkamp, W.: Twitter hashtags: Joint Translation and Clustering, pp. 1–3 (2011)
15. Sourin, A.: Computer Graphics: From a Small Formula to Cyberworlds. Prentice-Hall Inc., Singapore (2006)
16. Jorgensen, Z., Yu, T.: On mouse dynamics as a behavioral biometric for authentication. In: Proceedings of 6th ACM Symposium on Information, Computer and Communications Security, pp. 476–482 (2011)
17. Bours, P.: Continuous keystroke dynamics: A different perspective towards biometric evaluation. Inf. Secur. Tech. Rep. **17**(1), 36–43 (2012)
18. Frank, M., Biedert, R., Ma, E., Martinovic, I., Song, D.: Touchalytics: on the applicability of touchscreen input as a behavioral biometric for continuous authentication. IEEE Trans. Inf. Forensics Secur. **8**(1), 136–148 (2013)
19. Guo, Y., Yang, L., Ding, X., Han, J., Liu, Y.: OpenSesame: unlocking smart phone through handshaking biometrics. In: Proceedings of IEEE INFOCOM, pp. 365–369. IEEE, April 2013
20. Feng, T., Zhao, X., Shi, W.: Investigating mobile device picking-up motion as a novel biometric modality. In: Proceedings of IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS), pp. 1–6 (2013)
21. Jiang, W., Xiang, J., Liu, L., Zha, D., Wang, L.: From mini house game to hobby-driven behavioral biometrics-based password. In: Proceedings of 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), pp. 712–719. IEEE, July 2013
22. Olejnik, L., Castelluccia, C.: Towards web-based biometric systems using personal browsing interests. In: Proceedings of Eighth International Conference on Availability, Reliability and Security (ARES), pp. 274–280. IEEE, September 2013
23. Fridman, A., Stolerman, A., Acharya, S., Brennan, P., Juola, P., Greenstadt, R., Kam, M.: Decision fusion for multi-modal active authentication. IT Prof. **15**(4), 29–33 (2013)
24. Bailey, K.O., Okolica, J.S., Peterson, G.L.: User identification and authentication using multi-modal behavioral biometrics. Comput. Secur. **43**, 77–89 (2014)
25. Cohen, W., Ravikumar, P., Fienberg, S.: A comparison of string metrics for matching names and records. In: Proceedings of KDD Workshop on Data Cleaning and Object Consolidation, vol. 3, pp. 73–78, August 2003
26. Ross, A., Jain, A.: Information fusion in biometrics. Pattern Recogn. Lett. **24**(13), 2115–2125 (2003)