



Online User Footprints Prediction using Naïve Bayes, k Nearest Neighbor and Random Forest

Sanjeev Dhawan¹, Kulvinder Singh², Vijender Kumar³

^{1,2}Faculty of Computer Science & Engineering, ³PG Student of M.Tech. (Computer Engineering),

^{1,2,3}Department of Computer Science and Engineering,

University Institute of Engineering and Technology,

Kurukshetra University, Kurukshetra-136119, Haryana, India.

Abstract: Today the popularity of online social media is going on increasing. People want to use various services of social sites and hence they have multiple accounts across various social websites. The users of these social websites have to provide their basic personal information. The public availability of this information will do a favour for online attackers. This revealing of personal data may lead to compromise with security and privacy issues of online user. The attacker can extract relevant information about a user and may use it illegally for profile cloning, online phishing, privacy threats etc. In this paper some methodologies have proposed to overcome these problems. A concept called as "Online Digital Footprints" is introduced in proposed work. An online digital footprint is a collection of user's information gathered from multiple social websites. The datasets are collected manually from Facebook and Orkut. A number of similarity checking techniques are used to find similarity score between two features of two different social networks accounts. Thereafter, a feature matrix is generated based on the results given by Smith Waterman technique. In the present paper an attempt has been to find accuracy by implementing various classification techniques like k-nearest neighbour (kNN), Random forest (RF), Naïve Bayes (NB) and finally to produce results by using similarity techniques like recall, precision and f-measure.

Keywords: cross-folksonomy, profile cloning, similarity techniques, social network, user profile.

I. Introduction

At present, every person is living in two different worlds: one is digital world and another is real world. This digital world is going on more and more important than our real life. The reason is that people are so busy that they have no time to meet their near and dear ones and majority of times they are connected with each other through social networking sites. Nowadays the people cannot imagine their life without social networks. Only Facebook and MySpace both have over 250,000,000 accounts and still it is growing rapidly. The users share their personal information on social site. This became a treasure for online attackers. The attackers can easily collect the personal information and then use it for making a clone profile. Therefore there is an urgent need to design some techniques which can be used to detect and prevent cloning. One more problem there is that some users do not provide the complete and unambiguous personal information. For example some users create their accounts on different social sites with different user names or date of birth. These are the big challenges in front of researchers. These challenges can be resolved efficiently by a concept called as "Online User Footprint". The question is that what is this online user footprint? This is nothing but a digital identity of a user on the social networking sites. With the growing popularity and usage of online social media services, people now have several accounts on multiple and diverse services like Facebook, LinkedIn, Twitter and YouTube. The online digital footprints are created from publically available information. Personalization, profile management, detecting malicious behavior of users, and detecting cloned accounts can be done by using digital footprints. Personalization is nothing but authentication of identity of that user. In profile management, the user can manage his account's privacy and security parameters. The most important advantage of analyzing user's online digital footprints is to protect users from potential privacy and security risks which arises due to the huge publicly available user information. The users share different type of information on these sites. But when this different type of information is combined from these different social networking sites, then user's online social footprints are formed. In other words, it is the unique online identity of an online user by which that user is identified on different social networking sites. To create user footprint, linking of profiles across multiple social sites is necessary. To link these profiles we have to compare the information provided by the user. According to the nature of the information different techniques are used for comparison. For example, for collating usernames of two profiles SVM, kNN, Naïve Bayes etc. can be used. But for comparing display images of two different profiles PSNR and MSE are calculated. As shown in figure 1, a person named Michael Brown is available on Orkut, Facebook, and MySpace. Each social site asks for different type of information. All these information is gathered to create user footprint.

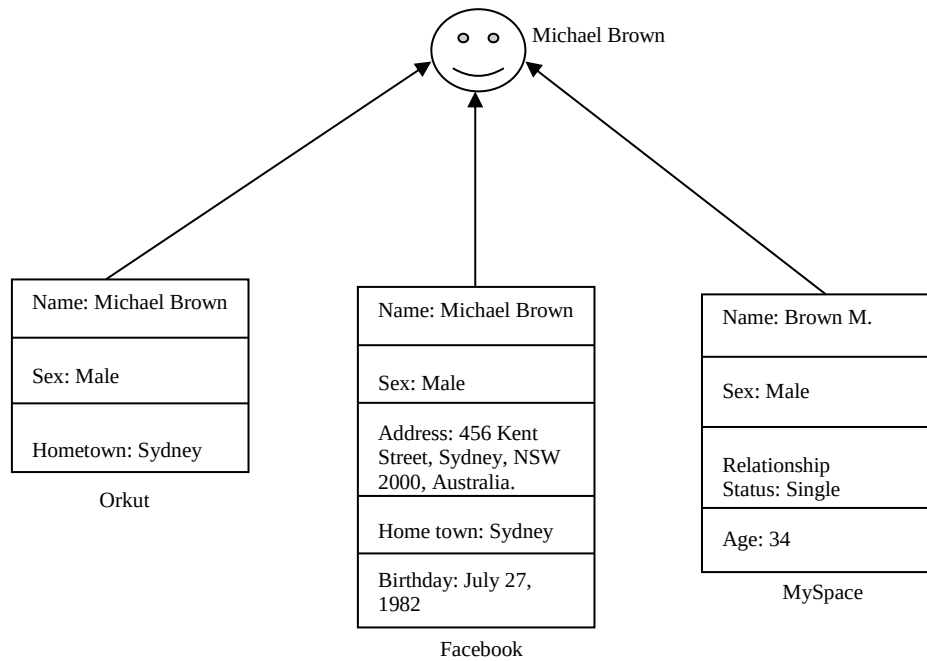


Figure 1: Creating digital footprint

II. Related Work

To link social networks on web, Jennifer Golbeck [1] in 2008 preferred the concept of Semantic Web. Semantic Web was used to store the social networking data at distributed locations and user's profile was represented in friend of a friend format (FOAF). The different social websites used FOAF format to store the data of their users and identify each user by FOAF profile. After making FOAF i.e. friend of a friend profiles, different accounts of same user on different websites were merged to represent single user. The Inverse Functional Property of Web Ontology Language OWL was used as important semantic feature to connect an online user with a unique identifier. Every single user on different social websites was connected to form unified graphs. In similar way Martin Szomszor *et al.* [2, 4] used the concept of folksonomies to pick out the correlated user profiles which tag the similar resources. From multiple folksonomies, they compared tag clouds of different users to find and align the user's separate identities. This work gave a good contribution to find similar user profiles over multiple communities. Through their work they shown that the filtering using cross-folksonomies increase the alignment between users tag clouds but to a limited extend. In their next work, they did a work to determine the semantic modeling of user interests based on cross folksonomy analyses. They showed that when multiple tag clouds of multiple users are combined then a vast amount of interest profile could be generated. In 2010 Tereza Iofciu *et al.* [5] used the social tagging system for user identification across communities. They took two entities for identification, first is username and second one is tag. Different approaches and techniques were implemented to measure the distance between two profiles and an accuracy of 60% to 80% was achieved. They examined user profiles from three different social tagging services: Flickr, Delicious and StumbleUpon. A forward step in the same field was taken by Jan Voscky *et al.* [6] in 2009. They proposed an approach for identification of users across multiple social websites. A method is proposed to identify users based on profile matching. In their proposed work, the Web profiles from two different social networking websites were studied and fields in the profiles which are suitable for the profile-comparison process were taken into account. Similarly in 2011, to link user profiles across multiple websites, Daniele Perito *et al.* [7] judged the feasibility of using the usernames to connect profiles of same user across communities. The assumption is that the probability that two usernames refer to the same physical person strongly depends on the entropy of the username string itself. The experiment showed that a significant portion of the users' profiles could be linked using their username. In 2005, an approach regarding the privacy of personal information of an online user was proposed by Ralph Gross [8]. The experiments were carried out on Facebook. The patterns of information exposed and privacy conditions were defined. The privacy settings of profiles and the amount of information exposed were studied. They highlighted potential attacks on various aspects of their privacy, and showed that only a minimal percentage of users change the highly permeable privacy preferences. In 2009, Joseph Bonneau *et al.* [9] proposed some methods to extract the user information from social networking sites. Their main motive was to check the ease to which third parties can access the user information from OSNs. Experiments were carried out on Facebook. Two results came out. First was, several methods were described to extract the user data. Secondly, they measured the efficiency of these methods. Then onwards in 2009, Leyla Bilge *et al.* [10] proposed their work to investigate that how easy it is for online attackers

to launch automated crawling and identity theft attacks against various social networking sites. Similarly in 2011, Georgios Kontaxis *et al.* [11] proposed his work to detect cloned profiles across multiple social websites. The architectural design and implementation details of a prototype system were presented that can be employed by users to investigate whether they have fallen victims to such an attack. Their results from the use of this prototype system proved its efficiency and also demonstrated its simplicity in terms of deployment by everyday users. After that in 2013, the same concept of clone detection was introduced by Mohammad Reza Khayyambashi [12]. At first, profile cloning and identity theft attack were introduced, and then a framework for detection suspicious identity was proposed. This approach was based on similarity of attribute and friend network. According to similarity measures which were computed in each step and by having predetermined threshold, it will be decided that which profile is clone which one is genuine. In next year i.e. in 2014, the detection and prevention of cloned profiles was done by Piotr Brodka *et al.* [13]. Two sophisticated methods are proposed of profile cloning detection and state-of-the-art research was also presented. The first method was based on the similarity of attributes from both profiles and the second method was based on the similarity of relationship networks. The methods were further evaluated with experiments and the results clearly described that the proposed methods were useful and efficient compared to existing methods. The paper also stressed that profile cloning in Facebook was not only possible but also fairly easy to perform. One step was taken forward in 2015 by M.A. Devmane [14], who proposed his work to prevent and detect cloned profiles across various communities. Some techniques are developed and applied to three social networking sites, Facebook, twitter and Google+ to detect the cloned profiles. The profile cloning and types of profile cloning was discussed. The mechanisms were designed to detect the same site profile cloning as well as cross-site profile cloning. In 2009, Danesh Irani *et al.* [15] proposed their work to create the large online digital footprints. Since the attackers can generate the footprints of a user from a single pseudonym like name, date of birth etc. So they worked for creating large digital footprints. They studied large online social footprints by collecting data on 13,990 active users. After parsing data from 10 of the 15 most popular social networking sites, they found that a user with one social network reveals an average of 4.3 personal information fields. For users with over 8 social networks, this average increases to 8.25 fields. Then onwards, Anshu Malhotra *et al.* [16] studied deeply about online user footprints and gave results, in 2012. Social Graph API, FriendFeed, and Profilactic were used to extract information about users from their individual identities from multiple social networking sites. They maintained their own dataset to create the digital footprints of the users. They used username, display name, description, location, profile picture, and number of connections to generate the digital footprints of the user. They applied context specific techniques (e.g. Jaro Winkler similarity, Wordnet based ontologies) to measure the similarity of the user profiles on different social networks. Then Sanjeev Dhawan *et al.* [17] in 2016 proposed their work in which they used random forest, Naïve Bayes and kNN for classification. They showed that random forest gave better accuracy, precision, f-measure and recall value.

III. Features of Online User Footprints

The online footprints can be created by collecting user information from different social sites. The user profile can be assumed a one dimensional array of user's information such as username, location, date of birth etc. Now this information can be used for profile matching on different social sites. From observations it was found that 4 to 8 pieces of information is asked by social sites. Here six features will be discussed, which are userid, display name, description, location, no. of connections and profile image. Also the different techniques for comparing respected features of two user profiles will be discussed.

A. *userid*: The userid is an alphanumeric string which identifies a user uniquely across multiple social sites. This is an important entity to authenticate a user on social sites. It may be possible that a single user has different usernames for different social sites. Various similarity available there to compare two usernames. *Jaro Winkler* is string matching technique which can be used to give a comparing score in-between 0 and 1. Also we can adopt *Jaccard* method.

B. *Display name*: It is an alphabetic string by which a user is called on social networking site. Display name consists of first name and last name of the user. A user can use different display name on different social sites. Here also *Jaro Winkler* and *Jaccard* method can be adopted to find similarity between two display names.

C. *Description*: Description is an alphanumeric string. This is a short story provided by user about him. We will consider two methods here. First method is *td-idf* method. Two steps are there. In first step all punctuations, full stops are removed. All alphabets are converted into lower case letter. Then a token set is prepared. In second step the cosine similarity method is used. Two token sets are compared to find the score between 0 and 1. Also we can adopt *Jaccard* and *Jaro Winkler* similarity techniques.

D. *Location*: The next profile feature that we will compare is location of the user. The users also provide their current location or their living city on social sites. Now there are two cases. Token set is generated by removing punctuations and full stop. The alphabets are converted into lower case letter. Then *Jaccard* and *Jaro Winkler* can be adopted to find similarity score between 0 and 1.

E. *Number of Connections*: It is purely a numeric value. For different social sites the number of connections can be defined in different way. For example, for twitter the number of connections for a user is number of people

he will follow. In case of facebook, the number of connections is the number of people with whom the user has friendship. Jaccard or Jaro Winkler can be used for comparing number of connections.

F. *Profile Image*: Comparison of two profile images that they belong to same person or not, is difficult task. Jaccard or Jaro Winkler techniques cannot be used for image comparison. Two methods are available there. Calculating MSE and PSNR. First of all the images are converted into grey scale and resized. Then for each pixel, corresponding value is stored at right place in $m \times n$ matrix. Then noise and signal value is calculated by doing some operations on two matrices. After that MSE is calculated from noise and signal to find comparison factor between two images.

IV. Experimental Evaluation & Problem Implementation

Java-ML library and weka are used for implementation. The aim is to interlink the profiles of same user across multiple social networking sites. This interlinking can be done by comparing the similar information provided by user on two different social sites. Various similarity techniques are available here for comparing the similar features. SmithWaterman is preferred because its accuracy is maximum among others. Three different classification techniques are used: Bayes, Random Forest and k-nearest neighbor. At last, the accuracy, precision, F measure, Recall and time taken by these three different techniques will be compared. Two different datasets are created by collecting data from Facebook and Orkut. First dataset is *Book1* and second dataset is *Book2*. Each dataset has six features < *userid*, *display name*, *location*, *description*, *no. of connections*, *profile image*>. For a particular profile, each feature is compared with corresponding features in all other profiles using SmithWaterman technique. The Smith–Waterman algorithm performs local sequence alignment; that is, for determining similar regions between two strings. Instead of looking at the total sequence, the Smith–Waterman algorithm compares segments of all possible lengths and optimizes the similarity measure. The algorithm was first proposed by Temple F. Smith and Michael S. Waterman in 1981. Smith–Waterman is a dynamic programming algorithm. As such, it has the desirable property that it is guaranteed to find the optimal local alignment with respect to the scoring system being used (which includes the substitution matrix and the gap-scoring scheme).

A matrix H is build as follows:

$$H(i,0) = 0, 0 \leq i \leq m;$$

$$H(0,j) = 0, 0 \leq j \leq n;$$

$$H(i,j) = \max \begin{cases} 0 \\ H(i-1,j-1) + s(a_i,b_j) & \text{Match/Mismatch} \\ \max_{k \geq 1} \{H(i-k,j) + W_k\} & \text{Deletion} \\ \max_{l \geq 1} \{H(i,j-l) + W_l\} & \text{Insertion} \end{cases}, 1 \leq i \leq m, 1 \leq j \leq n$$

Where:

- a, b = Strings over given alphabets;
- $m = \text{length}(a)$;
- $n = \text{length}(b)$;
- $s(a,b)$ is a similarity function;
- $H(i,j)$ = is the maximum Similarity-Score between a suffix of $a[1...i]$ and a suffix of $b[1...j]$;
- W_i is gap-scoring scheme;

Example:

- Sequence 1 = ACACACTA
- Sequence 2 = AGCACACA
- $s(a,b) = +2$ if $a=b$ (match), -1 if $a \neq b$ (mismatch);
- $W_i = -1$;

$$H = \begin{pmatrix} - & A & C & A & C & A & C & T & A \\ - & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ A & 0 & 2 & 1 & 2 & 1 & 2 & 1 & 1 \\ G & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ C & 0 & 1 & 3 & 2 & 3 & 2 & 3 & 2 \\ A & 0 & 2 & 2 & 5 & 4 & 5 & 4 & 4 \\ C & 0 & 1 & 4 & 4 & 7 & 6 & 7 & 6 \\ A & 0 & 2 & 3 & 6 & 6 & 9 & 8 & 8 \\ C & 0 & 1 & 4 & 5 & 8 & 8 & 11 & 10 \\ A & 0 & 2 & 3 & 6 & 7 & 10 & 10 & 12 \end{pmatrix} \quad T = \begin{pmatrix} - & A & C & A & C & A & C & T & A \\ - & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ A & 0 & \swarrow & \leftarrow & \nwarrow & \nwarrow & \nwarrow & \nwarrow & \nwarrow \\ G & 0 & \uparrow & \nwarrow & \uparrow & \nwarrow & \uparrow & \nwarrow & \uparrow \\ C & 0 & \uparrow & \nwarrow & \nwarrow & \nwarrow & \nwarrow & \nwarrow & \nwarrow \\ A & 0 & \nwarrow & \uparrow & \nwarrow & \nwarrow & \nwarrow & \nwarrow & \nwarrow \\ C & 0 & \uparrow & \nwarrow & \uparrow & \nwarrow & \nwarrow & \nwarrow & \nwarrow \\ A & 0 & \nwarrow & \uparrow & \nwarrow & \uparrow & \nwarrow & \nwarrow & \nwarrow \\ C & 0 & \uparrow & \nwarrow & \uparrow & \nwarrow & \uparrow & \nwarrow & \nwarrow \\ A & 0 & \nwarrow & \uparrow & \nwarrow & \uparrow & \nwarrow & \uparrow & \nwarrow \end{pmatrix}$$

To obtain the optimum local alignment, start with the highest value in the matrix (i,j) . Then, go backwards to one of positions $(i-1,j)$, $(i,j-1)$, and $(i-1,j-1)$ depending on the direction of movement used to construct the matrix. This methodology is maintained until a matrix cell with zero value is reached. In the example, the highest

value corresponds to the cell in position (8,8). The walk back corresponds to (8,8), (7,7), (7,6), (6,5), (5,4), (4,3), (3,2), (2,1), (1,1), and (0,0). Once finished, the alignment is reconstructed as follows: Starting with the last value, reach (i,j) using the previously calculated path. A diagonal jump implies there is an alignment (either a match or a mismatch). A top-down jump implies there is a deletion. A left-right jump implies there is an insertion. For the example, the results are:

Sequence 1 = A-CACACTA

Sequence 2 = AGCACAC-A

A feature matrix is generated which store the comparison score in between 0 and 1. After that the classification techniques are implemented on this feature matrix to find Accuracy, Precision, Recall, F-Measure. These four terms are defined as: In statistical analysis of binary classification, the F_1 score (also F-score or F-measure) is a measure of a test's accuracy. It considers both the precision p and the recall r of the test to compute the score: p is the number of correct positive results divided by the number of all positive results, and r is the number of correct positive results divided by the number of positive results that should have been returned. The F_1 score can be interpreted as a weighted average of the precision and recall, where an F_1 score reaches its best value at 1 and worst at 0. The traditional F-measure or balanced F-score (F_1 score) is the harmonic mean of precision and recall:

$$F_1 = 2 * \left(\frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \right)$$

In pattern recognition and information retrieval with binary classification, precision (also called positive predictive value) is the fraction of retrieved instances that are relevant

$$\text{Precision} = \frac{|\{\text{Relevant documents}\} \cap \{\text{Retrieved documents}\}|}{|\{\text{Retrieved documents}\}|}$$

Recall (also known as sensitivity) is the fraction of relevant instances that are retrieved. Both precision and recall are therefore based on an understanding and measure of relevance.

$$\text{Recall} = \frac{|\{\text{Relevant documents}\} \cap \{\text{Retrieved documents}\}|}{|\{\text{Relevant documents}\}|}$$

The accuracy is the proportion of true results (both true positives and true negatives) among the total number of cases examined.

As shown in flow chart, first of all data is read from Facebook and Twitter. After that the data is preprocessed. In this phase commas, inverted commas are removed. Capital letters are converted into lower case letters. In the third phase the similarity techniques are applied to find the similarity score. The feature matrix is obtained from the similarity scores produced. In fourth phase the classifiers are applied to the feature matrix obtained in third phase. In the fifth phase performance of different classifiers used are compared on the bases of accuracy, precision, recall and f-measure. The comparison table was made. It is clear from the table that Accuracy, Precision, Recall and F-Measure of Random Forest is maximum. Whereas the time taken to make comparison is more than Naïve Bayes. So overall result is that random forest gives better performance than Naïve Bayes and kNN. The graphical representation of classifiers based on different features.

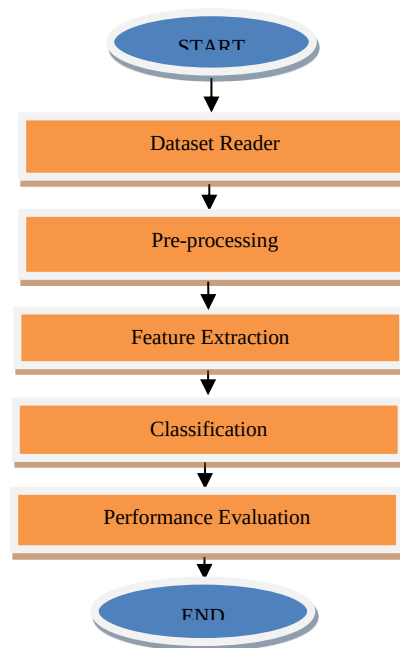


Figure 2: Flow chart of proposed work

Table 1: Comparison of various classifiers used

	Accuracy	Precision	Recall	F-Measure	Time taken
NB	0.992753623	0.973365617	0.933837429	0.952658662	136
RF	1	1	1	1	307
KNN	0.956521739	0.479128857	0.49905482	0.488888889	428

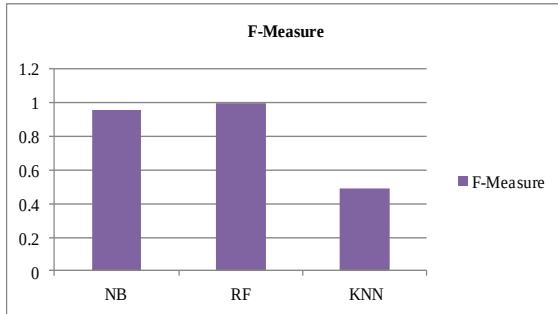


Figure 3: F-Measure calculation

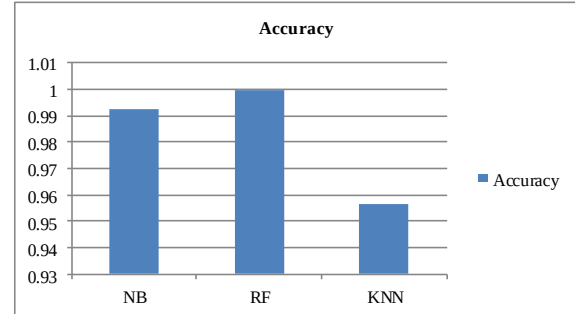


Figure 6: Accuracy calculation

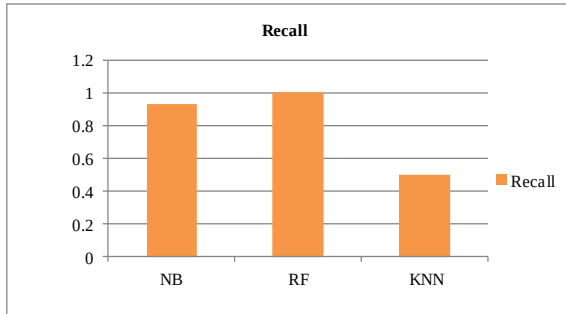


Figure 4: Recall calculation

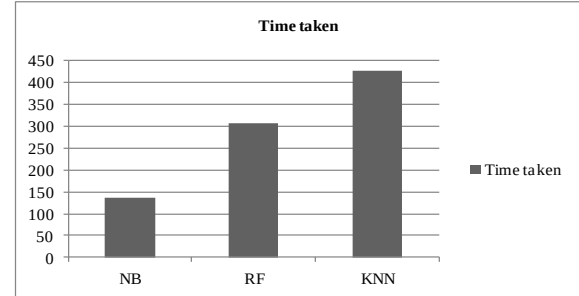


Figure 7: Time taken calculation

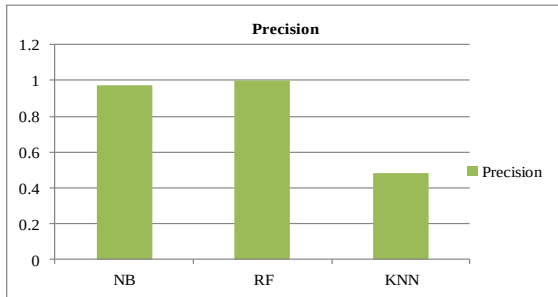


Figure 5: Precision calculation

V. Conclusion

The digital footprints, profile cloning and linking of user profiles, all are related terms. The reason is that to create users' footprints, the linking of various profiles of a single user on multiple social sites is required. Also the profile cloning can be detected by linking the users' profiles. For profile matching, various similarity measuring techniques are used such as Jaro Winkler, Jaccard, TD-IDF etc. Further classifiers were used such as SVM, Naïve Bayes, Random Forest, kNN for finding accuracy, precision, recall and F_1 . In future weighted kNN, genetic algorithm, perceptron can also be used for better performance and accuracy.

References

- [1] J. Golbeck and M. Rothstein, "Linking social networks on the web with foaf: a semantic web case study," ser. AAAI'08, 2008, pp. 1138–1143.
- [2] M. Szomszor, H. Alani, I. Cantador, K. O'Hara, and N. Shadbolt, "Semantic modelling of user interests based on cross-folksonomy analysis," ser. ISWC '08, 2008, pp. 632–648.
- [3] T. Vander Wal. Folksonomy definition and Wikipedia, November, 2005.
- [4] M. N. Szomszor, I. Cantador, and H. Alani, "Correlating user profiles from multiple folksonomies," ser. HT '08, 2008, pp. 33–42.
- [5] T. Iofciu, P. Fankhauser, F. Abel, and K. Bischoff, "Identifying users across social tagging systems," 2011. Available: www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2779.
- [6] J. Vosecky, D. Hong, and V. Y. Shen, "User identification across multiple social networks," in NDT'09, July 2009.
- [7] D. Perito, C. Castelluccia, M. A. Kaafar, and P. Manils, "How unique and traceable are usernames?" in PETS, 2011, pp. 1–17.

- [8] R. Gross and A. Acquisti, "Information revelation and privacy in online social networks," in WPES '05 Proceedings of the ACM workshop on Privacy in the electronic society.
- [9] J. Bonneau, J. Anderson, F. Stajano, and R. Anderson, "Eight Friends Are Enough: Social Graph Approximation via Public Listings," in SNS '09: Proceeding of the 2nd ACM Workshop on Social Network Systems, 2009.
- [10] L. Bilge, T. Strufe, D. Balzarotti, and E. Kirda, "All your contacts are belong to us: automated identity theft attacks on social networks."
- [11] G. Kontaxis, I. Polakis, S. Ioannidis, and E. Markatos, "Detecting social network profile cloning," in PerCom, March 2011, pp. 295–300.
- [12] Mohammad Reza Khayyambashi, Fatemeh Salehi Rizi, "An approach for detecting profile cloning in online social networks" in 7th international conference, IEEE, 2013.
- [13] Piotr Brodka, Mateusz Sobas, Henric Johnson, "Profile Cloning Detection in Social Networks" European Network Intelligence Conference, 2014.
- [14] M.A. Devmane, Dr. N.K. Rana, "Detection and Prevention of Profile Cloning in Online Social Networks" IEEE, 2014.
- [15] D. Irani, S. Webb, K. Li, and C. Pu, "Large online social footprints—an emerging threat," in CSE '09, vol. 3, Aug. 2009, pp. 271 – 276.
- [16] Anshu Malhotra, L. Totti, W. Meira Jr., P. Kumaraguru, V. Almeida, "Studying User Footprints in Different Online Social Networks", IEEE , 2012.
- [17] Sanjeev Dhawan, Kulvinder Singh, Vijender Kumar, "Managing Online User Footprints", 2016, Vivechana: National Conference on Advances in Computer Science and Engineering (ACSE-2016).