International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 03 | Mar -2017 www.irjet.net p-ISSN: 2395-0072

A STUDY ON IDENTIFYING SIMILAR USERS ACROSS MULTIPLE SOCIAL MEDIA SITES

Anju Viswam[1], Gopu Darsan[2]

[1] *PG scholar, Department of Computer Science and Engineering, Sree Buddha College of Engineering, Alappuzha, India.*
[2] *Assistant Professor, Department of Computer Science and Engineering, Sree Buddha College of Engineering, Alappuzha, India.*

Abstract - *Within this course of time, social media sites have gained a great attention. People rely on social media for different purposes. Not all the online social media will provide the same service. For this reason people tend to have multiple accounts on multiple social media sites. It is challenging and also interesting to identify the account that belongs to the same user in multiple social media sites. Many researchers have been conducted to match the user accounts in different social media sites. In this paper we have concentrated on different techniques used by the researchers to identify the similar accounts in multiple social media sites. Hence this paper will give an idea about the techniques that solves the problem of identifying user accounts across multiple online social media sites.*

***Key Words*: cross-communities, k-anonymity, l-diversity, social networks, social-tagging systems, user identification.**

1. INTRODUCTION

The users of internet are increasing and a large number of them are an active member of a social network. People rely on different social media networks for news, information and opinion of other people about different subjects. For example, people use Facebook for chat with their friends and families and also to share about their aspects of personal lives and Twitter to post about the things they are more passionate. For this reason people tend to have multiple accounts in multiple social media sites.

Discovering the same user accounts that belong to the same user is becoming a growing interest among researchers. Though it is more challenging, it is useful in developing many applications. It is useful to aggregate information about a single user. Merged information about a single user can give a detailed view about all available data. This information will be helpful to construct a complete social graph that helps in many applications such as information retrieval, collaborative filtering, sentiment analysis etc. It is also useful in modern marketing. As the modern marketing deals with targeting marketing with promotional messages, it is very useful to discover the same user accounts. Once the target customer is identified, the marketer does not have to bother the customer with multiple messages that has the same content. Identifying the same user accounts among multiple sites is also useful in the application automatic contacts' merging that happens almost in most of the mobile phones.

So identifying the same user accounts among multiple online social media sites is a challenging research area. Many studies where based on the profile attributes of the users, contents posted by the users and also by analyzing the network structures. Some of these studies are explained below.

2. LITERATURE SURVEY

As explained before the current studies depend on profile attributes, contents and network structures.

2.1 User identification based on profile

Username is a publicly available feature in the profile of the users. Perito et al. [1] presented an analytical model based on binary classifiers to calculate the similarity of usernames to identify the similar user accounts. An unsupervised approach was followed by Liu et al. [2] for linking users across multiple online sites. They computed the n-gram probabilities of the usernames to identify the rare and common usernames used by the persons. R. Zafarani and H. Liu [3] matched the users by extracting the usernames that appear in the URLs of the web pages. R. Zafarani and H. Liu [4] further developed a supervised learning approach to study the behaviors of various usernames chosen by the users among cross-communities. The usernames are publicly available and it can be faked by anyone. So these studies have some limitations.

Acquisti et al. [5] used the profile photos for matching the users. They used the face recognition algorithm and conducted the experiment on Facebook. Facebook profile photos are publicly available and it can be easily extracted. But this algorithm cannot be applied to large network because many users can use the same profile photos.

Iofciu et al.[6] linked the users based on the tags and user ids across social tagging systems. The tagging behaviors of the users where used to construct the user profiles based on the symmetric variant BM25 to link the users using the tags. Though it supports cross-platform, this technique cannot assure about the privacy. Motoyama and G. Varghese [7] used the classifier based on boosting to calculate the similarity of usernames. They analyzed the profile attributes such as the

name, geographical location, age etc. Boosting is a method to combine weak hypotheses to strong hypothesis. Goga [8] also used the profile attributes such as name, location and photos to find the similar accounts. He employed the Naïve Bayes classifier to calculate the probability of matching accounts. Elie Raad et al [9] used decision making classification algorithm to match the similar accounts based on the profile attributes. A weighted ontology-based user profile resolution technique was presented by Cortis [10] to discover the multiple profiles corresponding to the same user. A profile matching algorithm with Text Analytics was proposed to analyze the profile attributes in the profiles. The algorithm can be applied both in semi structured and structured profiles.

Fabien Abel et al. [11] studied the impact of cross-system user modeling in today's social media sites. The individual users on the online sites were analyzed and their characteristics where studied based on the profiles distributed on the web. The tagging behavior of the users was also analyzed. They used Mypes, a service that allows the aggregation of tag-based as well as the form-based profiles. Mypes includes the features like alignment, linkage and enrichment of various profiles of users distributed among different sites.

O.de Vel [12] proposed mining email content for author identification forensics. Email communication is becoming important with the emergence of the internet. Several cyber crimes were reported with the emails. It includes identity theft, plagiarism etc. All these occur when the identity of the author who is sending the email is not revealed. Here vector machine learning algorithm is used to identify the identity of the true author. For text categorization, many learning methods were proposed.

Vector-based comparison algorithm was used by J.Vosecky et al. [13] for identifying users on multiple sites based on profile matching. Here the profiles where represented as a vector consisting of individual profile attributes. Then the similarity score between the profiles in multiple sites where calculated. A weighting vector was utilized for the similarity score calculation as the profile will contain different kinds of information. If the similarity score reaches certain threshold, the two profiles will considered for the same individual. This method will be impractical when the profiles of users contain partial or missing datas.

2.2 User identification based on contents

Here the users are identified based on the time and locations the users post the contents and also with the writing style of the contents.

Zheng et al. [14] proposed a frame work to identify the authors based on their writing style on review sites. A privacy measurement was adapted to measure reviews to identify the anonymous reviews. The statistical tools such as Naïve Bayes Model and Kullback-Leiber Divergence Metric were used to find the anonymous reviews by anonymous users. The analysis of anonymous reviews was collected by extracting the tokens such as unigrams, digrams, rating and category from the reviews. These techniques can be adopted in review sites to get feedback about the reviews.

Xiangnan Kong et al. [15] proposed multi-network anchoring model for discovering multiple accounts of the same user distributed among multiple social media networks. The multi-network anchoring method will extract heterogeneous features from multiple social networks for the prediction of anchor links including users temporal, spatial, social and text information. They used a binary classifier on the training set data for the prediction of anchor links. The matching profiles were discovered based on the scores of the binary classifier. This method is effective for predicting anchor links with one-to-one relationships across multiple social media networks.

For correlating users across various sites Oana Goga et al. [16] examined the posted content of the users. They examined the writing style of the post, geo-location that is attached to the posts. They identified accounts from different sites that correspond to the same user. The user's activities were focused to collect the name or date of birth of the users. Mainly three types of features of posts were analyzed. The language, timing and location from where the content was posted are the features. Many sites will attach the geo-location with the users' content. All posts in multiple sites will have the identical timestamps. The language used by the users has characteristic writing style to identify the users. Based on this an attacker model was developed to identify similar accounts from multiple sites. As it can identify identical users among different sites, it poses some threats as it concentrates on the posted contents.

R.Zheng et al. [17] developed a framework for identification of author of online messages to address the identity-tracing problem. Here four types of writing-style features such as syntactic, structural, lexical and content-specific features where extracted from online messages by means of a feature extractor. Based on this inductive learning algorithms were used to build classification models to identify the true identity of authors who are posting the online messages.

2.3 User identification based on network structures

Identifying users across multiple social networks can be done by analyzing the network structures and seed. Though network based user identification is challenging many researches were based on network structures. Some of the studies are explained below.

Aravind Narayanan and Vitaly Shmatikov [18] presented a framework for privacy analyzing and anonymity in social

media networks. They developed a re-identification algorithm for finding anonymity in social networks. As online social networks share sensitive information about users privacy it should be protected by anonymization. Several de-anonymization attacks and different types of attackers who attack the online social media data were analyzed. A seed identification algorithm was developed for breaching privacy in online social media networks. A generic re-identification algorithm for anonymized social media networks was developed. This algorithm uses only the network structure and does not make any prior assumptions. But this technique can result in privacy breaches.

A joint link attribute (JLA) approach was introduced by Bartunov et al. [19] for finding multiple profiles of the same user in different social media networks. The user identity resolution problem was formulated in terms of Conditional Random Fields model constructed by means of a social network graph. The unwanted projections were removed by means of filtering based on a numerous network features to improve the results. The JLA model was developed based on the projections of the adjacent nodes in the constructed social graph and assumes that if the nodes are fully connected they should have lowest network distance. The algorithm is not robust for user identification problem.

An efficient reconciliation algorithm was proposed by Korula and Lattanzi [20] to identify the all accounts belonging to the same individual in multiple social networks. An efficient parallel algorithm to formalize the user identification problem mathematically was developed by them. A formal model was developed for the graph reconciliation problem that captures the trusted links to identify the users across different social media networks. The model can generate permutations of the graph and list of trusted links for some users across different social media networks. The parallel algorithm will solve the graph reconciliation problem. The algorithm is completely depending on the graph structure and the set of links across the network.

Bin Zhou and Jian Pei [21] present a solution to the neighborhood attacks. With some knowledge about the users, an adversary can attack the privacy of some users easily. They modeled a social network graph for finding the anonymization and privacy in social networks and also to get the adversary background knowledge. The k-anonymity method [22] was used to extract the neighborhoods of all vertices in the graph and proposed an efficient neighborhood component coding technique to represent the neighborhood of all vertices. But this method may still leak privacy.

The k-anonymity and l-diversity approaches were proposed by Zhou and Pie [22] for preserving privacy against neighborhood attacks in social media networks. Preserving privacy in social data becomes an important concern. They modeled k-anonymity and l-diversity model s for finding neighborhood attacks in social media networks. The k-anonymity and l-diversity models make assumptions about attackers and their knowledge. The k-anonymity model is to protect the privacy of vertices and kept the re-identification confidence lower than a pre-defined threshold. If a graph G is k-anonymous, then it has confidence lower than 1/k. An l-partition divides the vertices in the graph into equivalence groups of vertices such that frequency of vertices is less than or equal to1/l of the vertices associate with a sensitive label. Both the methods have some privacy problems.

3. CONCLUSIONS

Identifying same user accounts across multiple social media sites is interesting and also it is very challenging. We conducted a survey of different techniques that is used for identifying the similar user accounts in various social networking sites. Each technique can identify only a portion of similar accounts. Because the profile attributes can be easily faked by others as it is publicly available, contents and network structures are difficult to extract. So the technique can be chosen according to the human user.

REFERENCES

[1] D. Perito, C. Castelluccia, M. A. Kaafar, and P. Manils, "How unique and traceable are usernames?" in Proc. 11th Int. Conf. Privacy Enhancing Technol., 2011, pp. 1–17.

[2] J. Liu, F. Zhang, X. Song, Y. I. Song, C. Y. Lin, and H. W. Hon, "What's in a name?: An unsupervised approach to link users across communities," in Proc. 6th ACM Int. Conf. Web Search Data Mining, 2013, pp. 495–504.

[3] R. Zafarani and H. Liu, "Connecting corresponding identities across communities," in Proc. 3rd Int. ICWSM Conf., 2009, pp. 354–357.

[4] R. Zafarani and H. Liu, "Connecting users across social media sites: a behavioral-modeling approach," in Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2013, pp. 41–49.

[5] A. Acquisti, R. Gross, and F. Stutzman, "Privacy in the age of augmented reality," in Proc. Nat. Acad. Sci., 2011, pp. 36–53, Available: https://www.usenix.org/legacy/events/sec11/tech/slides/acquisti.pdf

[6] T. Iofciu, P. Fankhauser, F. Abel, and K. Bischoff, "Identifying users across social tagging systems," in Proc. 5th Int. AAAI Conf. Weblogs Social Media, 2011, pp. 522–525.

[7] M. Motoyama and G. Varghese, "I seek you: searching and matching individuals in social networks," in Proc. 11th Int. Workshop Web Inf. Data Manage., 2009, pp. 67–75.

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

[9] K. Cortis, S. Scerri, I. Rivera, and S. Handschuh, "An ontology based technique for online profile resolution," in Proc. 5th Int. Conf. Social Informat.,

2013, pp. 284–298.

[10] E. Raad, R. Chbeir, and A. Dipanda, "User profile matching in social networks," in Proc. 13th Int. Conf. Netw.-Based Inf. Syst., 2010, pp. 297–304.

[11] F. Abel, E. Herder, G. J. Houben, N. Henze, and D. Krause, "Cross system user modeling and personalization on the social web," User Model. User-Adapted Interaction, vol. 23, pp. 169–209, 2013.

[12] O. De Vel, A. Anderson, M. Corney, and G. Mohay, "Mining e-mail content for author identification forensics," ACM Sigmod Rec., vol. 30, no. 4, pp. 55–64, 2001.

[13] J. Vosecky, D. Hong, and V. Y. Shen, "User identification across multiple social networks," in Proc. 1st Int. Conf. Netw. Digital Technol., 2009, pp. 360–365.

[14] M. Almishari and G. Tsudik, "Exploring likability of user reviews," in Proc. 17th Eur. Symp. Res. Comput. Security, 2012, pp. 307–324.

[15] X. Kong, J. Zhang, and P. S. Yu, "Inferring anchor links across multiple heterogeneous social networks," in Proc. 22nd ACM Int. Conf. Inf. Knowl. Manage. 2013, pp. 179–188.

[16] O. Goga, H. Lei, S. H. K. Parthasarathi, G. Friedland, R. Sommer, and R. Teixeira, "Exploiting innocuous activity for correlating users across sites," in Proc. 22nd Int. Conf. World Wide Web, 2013, pp. 447–458.

[17] R. Zheng, J. Li, H. Chen, and Z. Huang, "A framework for authorship identification of online messages: Writing style features and classification techniques," J. Amer. Soc. Inf. Sci. Technol., vol. 57, no. 3, pp. 378–393, 2006.

[18] A. Narayanan and V. Shmatikov, "De-anonymizing social networks," in Proc. IEEE 30th Symp. Security Privacy, 2009, pp. 173–187.

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

[20] N. Korula and S. Lattanzi, "An efficient reconciliation algorithm for social networks," arXiv preprint arXiv: 1307.1690, 2013.

[21] B. Zhou and J. Pei, "Preserving privacy in social networks against neighborhood attacks," in Proc. 24th IEEE Int. Conf. Data Eng., 2008, pp. 506–515.

[22] B. Zhou and J. Pei, "The k-anonymity and l-diversity approaches for privacy preservation in social networks against neighborhood attacks," Knowl. Inf. Syst, vol. 28, no. 1, pp. 47–77, 2011.