Survey Paper on Detecting Unknown or Fake User Accounts on Different Microblogging and Social Media Networks

Balaji Lase[1], Darshan Biyani[2], Sagar Sawarkar[3], Vaibhav Umale[4]

[1]Student, Information Technology, SKN-SITS,LONAVALA, balajilase11@gmail.com
[2] Student, Information Technology, SKN-SITS,LONAVALA, biyanidarshan11@gmail.com
[3] Student, Information Technology, SKN-SITS,LONAVALA, sagarsawarkar1995@gmail.com
[4] Student, Information Technology, SKN-SITS,LONAVALA, umalevaibhav555@gmail.com

ABSTRACT

The last few years have witnessed the emergence and evolution of a vibrant research stream on a large variety of online Social Media Network (SMN) platforms. Recognizing anonymous, yet identical users among multiple SMNs is still an intractable problem. Clearly, cross-platform exploration may help solve many problems in social computing in both theory and applications. Since public profiles can be duplicated and easily impersonated by users with different purposes, most current user identification resolutions, which mainly focus on text mining of users' public profiles, are fragile. Some studies have attempted to match users based on the location and timing of user content as well as writing style. However, the locations are sparse in the majority of SMNs, and writing style is difficult to discern from the short sentences of leading SMNs such as S in a Microblog and Twitter. Moreover, since online SMNs are quite symmetric, existing user identification schemes based on network structure are not effective. The real-world friend cycle is highly individual and virtually no two users share a congruent friend cycle. Therefore, it is more accurate to use a friendship structure to analyse cross-platform SMNs. Since identical users tend to set up partial similar friendship structures in different SMNs, we proposed the Friend Relationship-Based User Identification (FRUI) algorithm. FRUI calculates a match degree for all candidate User Matched Pairs (UMPs), and only UMPs with top ranks are considered as identical users. We also developed two propositions to improve the efficiency of the algorithm. Results of extensive experiments demonstrate that FRUI performs much better than current network structure-based algorithms.

Keywords: *Cross-Platform, Social Media Network, Anonymous Identical Users, Friend Relationship, User Identification*

1. INTRODUCTION

In the last decade, many types of social networking sites have emerged and contributed immensely to large volumes of real-world data on social behaviors. Twitter 1, the largest microblog service, has more than 600 million users and produces upwards of 340 million tweets per day [1]. Sina Microblog2, the primary Twitter-style Chinese mi-croblog website, has more than500 million accounts and generates well over 100 million tweets per day [2].

Due to this diversity of online social media networks (SMNs), people tend to use different SMNs for different purposes. For instance, RenRen 3, a Facebook-style but autonomous SMN, is used in China for blogs, while Sina Microblog is used to share statuses (Fig.1). In other words, every existent SMN satisfies some user needs. In terms of SMN management, matching anonymous users across different SMN platforms can provide integrated details on each user and inform corresponding regulations, such as targeting services provisions. In theory, the cross-platform explorations allow a bird's eye view of SMN user behaviors. However, nearly all recent SMN-based studies focus on a single SMN platform, yielding incomplete data. Therefore, this study investigates the strategy of crossing multiple SMN platforms to paint a comprehensive picture of these behaviors.

Nonetheless, cross-platform research faces numerous challenges. As shown in Fig.1, with the growth of SMN platforms on the Internet, the cross-platform approach has merged various SMN platforms to create richer raw data and more complete SMNs for social computing tasks. SMN users form the natural bridges for these SMN

platforms. The primary topic for cross-platform SMN research is user identification for different SMNs. Exploration of this topic lays a foundation for further cross-platform SMN research.

2. LITERATURE SURVEY

2.1 How unique and traceable are user nemeses?

AUTHORS: D. Perito, C. Castelluccia, M.A. Kaafar, and P. Manils.

This paper explores the possibility of linking users profiles only by looking at their usernames. The intuition is that the probability that two usernames refer to the same physical person strongly depends on the "entropy" of the username string itself. Our experiments, based on crawls of real web services, show that a significant portion of the users' profiles can be linked using their usernames. To the best of our knowledge, this is the first time that usernames are considered as a source of information when profiling users on the Internet.

2.2 Connecting corresponding identities across communities.

AUTHORS: R. Zafarani and H. Liu.

One of the most interesting challenges in the area of social computing and social media analysis is the so-called community analysis. A well known barrier in cross-community (multiple website) analysis is the disconnectedness of these websites. In this paper, our aim is to provide evidence on the existence of a mapping among identities across multiple communities, providing a method for connecting these websites. Our studies have shown that simple, yet effective approaches, which leverage social media's collective patterns can be utilized to find such a mapping. The employed methods successfully reveal this mapping with 66% accuracy.

2.3 Connecting users across social media sites: a behavioral-modeling approach.

AUTHORS: R. Zafarani and H. Liu.

Social media is playing an important role in our daily life. People usually hold various identities on different social media sites. User-contributed Web data contains diverse information which reflects individual interests, political opinions and other behaviors. To integrate these behaviors information, it is of value to identify users across social media sites. This paper focuses on the challenge of identifying unknown users across different social media sites. A method to relate user's identities across social media sites by mining users' behavior information and features is introduced. The method has two key components. The first component distinguishes different users by analyzing their common social network behaviors and finding strong opposing characters. The second component constructs a model of behavior features that helps to obtain the difference of users across social media sites. The method is evaluated through two experiments on Twitter and Sina Weibo. The results of experiments show that the method is effective.

2.4 Privacy in the age of aug-mented reality.

AUTHORS: A. Acquisti, R. Gross and F. Stutzman.

We investigate the feasibility of combining publicly available Web 2.0 data with off-the-shelf face recognition software for the purpose of large-scale, automated individual re-identification. Two experiments illustrate the ability of identifying strangers online (on a dating site where individuals protect their identities by using pseudonyms) and offline (in a public space), based on photos made publicly available on a social network site. A third proof-of-concept experiment illustrates the ability of inferring strangers' personal or sensitive information (their interests and Social Security numbers) from their faces, by combining face recognition, data mining algorithms, and statistical re-identification techniques. The results highlight the implications of the convergence of face recognition technology and increasing online self-disclosure, and the emergence of "personally predictable" information, or PPI. They raise questions about the future of privacy in an "augmented" reality world in which online and offline data will seamlessly blend.

2.5 I seek you: searching and matching individuals in social networks.

AUTHORS: M. Motoyama and G. Varghese

An online user joins multiple social networks in order to enjoy different services. On each joined social network, she creates an identity and constitutes its three major dimensions namely prole, content and connection network. She largely governs her identity formulation on any social network and therefore can manipulate multiple aspects of it. With no global identity to mark her presence uniquely in the online domain, her online identities remain unlinked, isolated and difficult to search.

Earlier research has explored the above mentioned dimensions, to search and link her multiple identities with an assumption that the considered dimensions have been least disturbed across her identities. However, majority of the approaches are restricted to exploitation of one or two dimensions. We make a rest attempt to deploy an integrated system Finding Name which uses all the three dimensions of an identity to search for a user on multiple social networks. The system exploits a known identity on one social network to search for her identities on other social networks. We test our system on two most popular and distinct social networks Twitter and Facebook. We show that the integrated system gives better accuracy than the individual algorithms. We report experimental endings in the paper.

3. EXISTING SYSTEM

The diversity of online social media networks (SMNs), people tend to use different SMNs for different purposes. For instance, RenRen 3, a Facebook-style but autonymous SMN, is used in China for blogs, while Sina Microblog is used to share statuses. In other words, every existent SMN satisfies some user needs.

3.1 DISADVANTAGES

3.1.1 Nearly all recent SMN-based studies focus on a single SMN platform, yielding incomplete data.
3.1.2 Many studies have addressed the user identification problem by examining public user profile attributes, including screen name, birth-day, location, gender, profile photo, etc.

4. PROPOSED SYSTEM

We proposed the FRUI algorithm. Since FRUI employs a unified friend relationship, it is apt to identify users from a heterogeneous network structure. Unlike existing algorithm,FRUI chooses candidate matching pairs from currently known identical users rather than unmapped ones. This operation reduces computational complexity, since only a very small portion of unmapped users are involved in each iteration. Moreover, since only mapped users are exploited, our solution is scalable and can be easily extended to online user identification applications. In contrast with current algorithms, FRUI requires no control parameters.

4.1 ADVANTAGES

4.1.1 Since only mapped users are exploited, our solution is scalable and can be easily extended to online user identification applications. In contrast with current algorithms.
4.1.2 Unlike existing algorithms, FRUI chooses candidate matching pairs from currently known identical users rather than unmapped ones. This operation reduces.

5. SYSTEM DESIGN

5.1 SYSTEM ARCHITECTURE

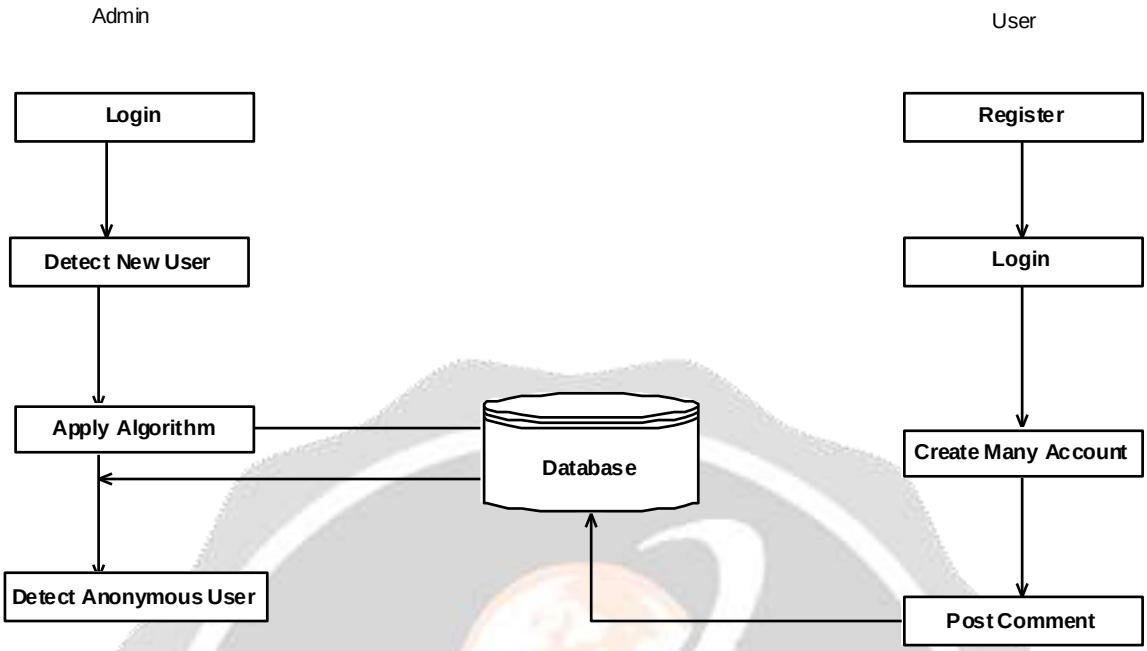

Fig-1: SYSTEM ARCHITECTURE

5.2 SYSTEM CONFIGURATION REQUIREMENTS

5.2.1. SOFTWARE REQUIREMENTS

- Operating System : Windows 7
- Technology : Java and J2EE
- Web Technologies : JSP, JavaScript, CSS
- IDE : Eclipse
- Web Server : Tomcat
- Database : My SQL
- Java Version : J2SDK1.7

5.2.1. HARDWARE REQUIREMENTS

- Hardware : Pentium Dual Core
- Speed : 2.80 GHz
- RAM : 1GB
- Hard Disk : 20 GB
- Floppy Drive : 1.44 MB
- Key Board : Standard Windows Keyboard
- Mouse : Two or Three Button Mouse
- Monitor : SVGA

5. CONCLUSION

This study addressed the problem of user identification across SMN platforms and offered an innovative solution. As a key aspect of SMN, network structure is of paramount importance and helps resolve de-anonymization user identification tasks. Therefore, we proposed a uniform network structure-based user identification solution. We also developed a novel friend relationship-based algorithm called FRUI. To improve the

efficiency of FRUI, we de-scribed two propositions and addressed the complexity. Finally, we verified our algorithm in both synthetic net-works and ground-truth networks.

Results of our empirical experiments reveal that net-work structure can accomplish important user identification work. Our FRUI algorithm is simple, yet efficient, and performed much better than NS, the existing state-of-art network structure-based user identification solution. In scenarios when raw text data is sparse, incomplete, or hard to obtain due to privacy settings, FRUI is extremely suitable for cross-platform tasks.

Moreover, our resolution can be easily applied to any SMNs with friendship networks, including Twitter, Face-book and Foursquare. It can also be extended to other studies in social computing with cross-platform problems such as targeted marketing [40], information retrieval [41], collaborative filtering [42], sentiment analysis [43] and more. In addition, since only the Adjacent Users are involved in each iteration process, our method is scalable and can be easily applied to large datasets and online user identification applications.

Identifying anonymous users across multiple SMNs is challenging work. Therefore, only a portion of identical users with different nicknames can be recognized with this method. This study built the foundation for further studies on this issue. Ultimately, it is our hope that a final approach can be developed to identify all identical users with different nicknames. Other user identification methods can be applied simultaneously to examine multiple SMN plat-forms. These methods are complementary and not mutually exclusive, since the final decision may rely on human user's involvement. Therefore, we suggest using these methods synergistically and considering strengths and weaknesses for the best results.

ACKNOWLEDGEMENT

This work was supported by the Fundamental Research Funds for the Central Universities, Research Funds of Renmin University of China (10XNI029), the Natural Science Foundation of Beijing under grant no. 4132067, the Natural Science Foundation of China under grant no. 71271211 and the Beijing Higher Education Young Elite Teacher Project under grant no. YETP1660.

The authors highly appreciated the valuable comments of the anonymous reviewers. They are also grateful to Chao Feng for his help on labeling the identical users.

REFERENCES

[1] Wikipedia, "Twitter, " http://en.wikipedia.org/wiki/Twitter. 2014.
[2] Xinhuanet, "Sina Microblog Achieves over 500 Million Users," http://news.xinhuanet.com/tech/2012-02/29/c_122769084.htm. 2014.
[3] D. Perito, C. Castelluccia, M.A. Kaafar, and P. Manils, "How unique and traceable are usernames?," Privacy Enhancing Technol-ogies(PETS'11), pp. 1-17, 2011.
[4] 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," Proc. of the 6thACM international conference on Web search and data mining(WDM'13), pp. 495-504, 2013.
[5] R. Zafarani and H. Liu, "Connecting corresponding identities across communities," Proc. of the 3rd International ICWSM Con-ference, pp. 354-357, 2009.
[6] R. Zafarani and H. Liu, "Connecting users across social media sites: a behavioral-modeling approach, " Proc. of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'13), pp.41-49, 2013.
[7] A. Acquisti, R. Gross and F. Stutzman, "Privacy in the age of aug-mented reality," Proc. National Academy of Sciences, 2011.
[8] T. Iofciu, P. Fankhauser, F. Abel, and K. Bischoff, "Identifying users across social tagging systems," Proc. of the 5th International AAAI Conference on Weblogs and Social Media, pp. 522-525, 2011.
[9] M. Motoyama and G. Varghese, "I seek you: searching and matching individuals in social networks," Proc. of the 11th inter-national workshop on Web Information and Data Management (WIDM'09), pp. 67-75, 2009.
[10] O. Goga, D. Perito, H. Lei, R. Teixeira, and R. Sommer, "Large-scale Correlation of Accounts across Social Networks," Tech-nical report, 2013.