ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 12, March 2016

Social Identity Linkage via Heterogeneous Behavior Modeling Using User Chatting Flow: A Survey

E. Emily
M.Phil scholar
PG & Research department of Computer Science
Government Arts college (Autonomous),
Salem-7
Tamilnadu - India
emilyeronimus@gmail.com

Mr. E. Jayabalan
Assistant Professor
PG & Research department of Computer Science
Government Arts college (Autonomous),
Salem-7
Tamilnadu - India
ejksrcas@gmail.com

Abstract: *Social identity linkage across different social media platforms is of critical importance to business intelligence by gaining from social data a deeper understanding and more accurate profiling of users. In this paper, we propose the model heterogeneous behavior by long-term topical distribution analysis and multi-resolution temporal behavior matching against high noise and information missing, and the behavior similarity are described by multi-dimensional similarity vector for each user pair, we build structure consistency models to maximize the structure and behavior consistency on users' core social structure across different platforms, thus the task of identity linkage can be performed on groups of users, which is beyond the individual level linkage in previous study. Thus we propose a WPM(Word per minute) formulation and calculating which is beneficial to distinguish between different users.*

Keywords**: Social identity linkage, structured Learning, heterogeneous behavior, multi-resolution temporal information matching, Chatting style.**

I. INTRODUCTION

The ability of assuming multiple identities has long been a dream for many people. Yet it is not until the late advent of online social networks that this ambition of millions has been made possible in cyber virtual world. In fact, the recent proliferation of social network services of all kinds has revolutionized our social life by providing everyone with the ease and fun of sharing various information's like never before (e.g., micro-blogs, images, videos, reviews, and location checkins). Meanwhile, probably the biggest and most intriguing question concerning all businesses is how to leverage this big social data for better business intelligence. In particular, people wonder how to gain thorough understanding of each individual user from the vast amount of online social data records. Unfortunately, information of a user from the current social scene is fragmented, inconsistent and disruptive. The key to unleashing the true power of social media is to link up all the data of the same user across different social platforms, offering the following benefits to user profiling.

Much of this work has been focused on more formal bodies of text. This fact allows us to attempt to use unique features and idiosyncrasies of people's informal online writing styles, in addition to traditional techniques and characteristics, to distinguish between writers. Being able to identify between writers in online chat has a number of significant uses. Court cases may often include a review of online chat communications, and this would be useful in potentially detecting fraudulent

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 12, March 2016

or tampered evidence. Furthermore, especially in a live chat conversation, this may also be used to identify social engineering attempts.

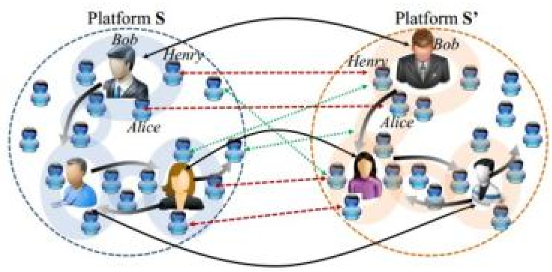

Fig.1: Identity linkage in various platforms

COMPLETENESS - Single social networks service offers only a partial view of a user from a particular perspective. Cross platform user linkage would enrich an otherwise-fragmented user profile to enable an all-around understanding of a user's interests and behavior patterns

CONSISTENCY - For various reasons, information provided by users on a social platform could be false, conflicting, missing and deceptive. Cross-checking among multiple platforms helps improve the consistency of user information.

UNRELIABLE ATTRIBUTES - How users register their names online varies among different platforms. For example, a user tends to add family name after "Adele" in English communities, and users are likely to put a Chinese name or bizarre characters before or after "Adele" for eccentricity in Chinese communities. To make things worse, people do not use their true names, women would not tell their true ages, and males even pretend to be females. Statistical models e.g. SVM or rule based models constructed with mere username and attribute analysis are far from being robust for accurate user linkage across online social communities.

CHATTING STYLE - we focus on the writing style of an individual, analyzing how it can be recognized given a portion of chat, and how personality traits come into play in this scenario. Two important results do emerge: some traits correlate significantly with some characteristics of the chatting style of people. Some of such features are very effective in recognizing a person among a gallery of diverse individuals. These facts seem to suggest that there are some personality traits that lead people to chat in a particular style, which turns out to be very recognizable.

II.LITERATURE REVIEW

1. **Structured Learning from Heterogeneous Behavior for Social Identity Linkage Siyuan Liu, Shuhui Wang, Feida Zhu** This paper aims to link user accounts across different social networks platforms. To deal with the challenges, we propose a framework, HYDRA, a multi-objective learning framework incorporating heterogeneous behavior model and core social networks structure. We evaluate HYDRA against the state-of the-art on two real data sets. Experimental results demonstrate hat HYDRA outperforms existing algorithms in identifying true user linkage across different platforms.

2. **Recognizing Chatting Style by Rohan Puttagunta, Nick Wu and Renjie** This work has been focused on more formal bodies of text. This fact allows us to attempt to use unique features and idiosyncrasies of people's informal online writing styles, in addition to traditional techniques and characteristics, to distinguish between writers. Being able to identify between writers in online chat has a number of significant uses. Court cases may often include a review of online chat communications, and this would be useful in potentially detecting fraudulent or tampered evidence. Furthermore, especially in a live chat conversation, this may also be used to identify social engineering attempts.

3. **Linking Personality, Style And Recognisability In Chats by Giorgio Roffo, Cinzia Giorgetta, Roberta Ferrario and Marco Cristani** Text chatting represents an hybrid type of communication, where textual information is delivered following turn-taking dynamics, which characterize spoken interactions. In this sense, it is interesting to understand whether social behaviour can emerge in chats, in the same way it characterizes face-to face exchanges.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 12, March 2016

4. **Connecting Users across Social Media Sites: A Behavioral-Modeling Approach by Reza Zafarani and Huan Liu Computer Science and Engineering Arizona State University** This paper aims to address the cross-media user identification problem. We introduce a methodology (MOBIUS) for finding a mapping among identities of individuals across social media sites. It consists of three key components: the rest component identities users' unique behavioral patterns that lead to information redundancies across sites; the second component constructs features that exploit information redundancies due to these behavioral patterns; and the third component employs machine learning for effective user identification. We formally deny the cross-media user identication problem and show that MOBIUS is erective indentifying users across social media sites.

5. **Link Prediction in Coupled Networks by Yuxiao Dong, Jing Zhang, Jie Tang, Nitesh V. Chawla, Bai Wang University of Notre Dame Notre Dame** Coupled network link prediction is different from the classical link prediction problem, which generally aims at predicting the future links in the next time period. Meanwhile, the proposed problem differs from link prediction in heterogeneous network, in which partial multi-typed links are given to predict the remaining single- or multi-typed links. Our problem is also different from the problem of transfer link prediction, which focuses on leveraging the estimated parameters in one network to improve the prediction performance of the other network based on the common features between the two networks. Finally, our problem is different from the problem of cross-domain link prediction, whereas it aims to predict the links in the cross network between two networks.

6. **Social Network Integration: Towards Constructing the Social Graph** In this work, we formulate the problem of social network integration. It takes multiple observed social networks as input and returns an integrated global social graph where each node corresponds to a real person. The key challenge for social network integration is to discover the correspondences or interlinks across different social networks.

III.MULTI OBJECTIVE STRUCTURE

Based on the heterogeneous behavior modeling from user attributes, we propose to learn the linkage function via a multi objective optimization framework

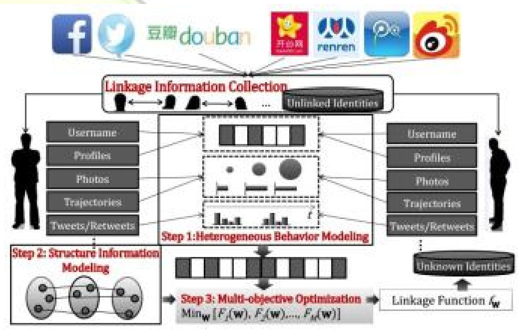

Fig 2 : Ambiguity between user on different social communities

Supervised Learning - Some social media platforms allow users to log in to different platforms with one account. For example, we can use a Facebook account to log in to Twitter. We collect such user-provided linkage information as the ground-truth label information. We notice that the labeled training pairs collected by our paradigm are much cleaner (precision over 95%) than the approach in (precision around 75%) where the labeled training pairs are automatically generated based on the uniqueness (n-gram probability) of user names. We also collect label information by user attribute matching as the pre-linked label information. By utilizing the collected label information, we minimize the structured loss on the labeled training data. Christo Ananth et al. [1] proposed a secure hash message authentication code. A secure hash message authentication code to avoid certificate revocation list checking is proposed for vehicular ad hoc networks (VANETs). The group signature scheme is widely used in VANETs for secure communication, the

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 12, March 2016

existing systems based on group signature scheme provides verification delay in certificate revocation list checking. In order to overcome this delay this paper uses a Hash message authentication code (HMAC). It is used to avoid time consuming CRL checking and it also ensures the integrity of messages. The Hash message authentication code and digital signature algorithm are used to make it more secure . In this scheme the group private keys are distributed by the roadside units (RSUs) and it also manages the vehicles in a localized manner. Finally, cooperative message authentication is used among entities, in which each vehicle only needs to verify a small number of messages, thus greatly alleviating the authentication burden.

Structure Consistency Modeling - We optimize the linkage function by maximizing both behavior similarity and social structure consistency between platforms. By constructing a positive semi definite second-order structure consistency matrix among candidate linked user pairs, our model is able to consider the global structure between platforms to identify the true linkages and filter out those false ones, as illustrated in Figure 3. Most importantly, it compensates for the shortage of ground truth linkage information for user-level supervised learning by propagating the linkage information along the core social structure (i.e., friends with the most frequent interactions) of each individual user.

Multi-objective Optimization - We learn the linkage function by jointly minimizing the two objective functions via a unified multi-objective optimization framework. We prove that our model is a generalized semi-supervised learning approach by leveraging both ground truth linkage information and social structure.

Behavior Similarity Modeling - We calculate similarity among pairs of users via heterogeneous behavior modelling.

Structure Information Modelling - We construct the structure consistency graph on user pairs by considering both the core network structure of the users and their behaviour similarities.

IV.CONCEPTUAL METHODOLOGY

The language style of a user including personalized wording and emotion adoption is usually well reflected in comments, tweets and re-tweets, which is beneficial to distinguishing between different users. To model a user's characteristic style, we extract the most unique words of each user by a simple term frequency analysis on the whole database

A. WORD STYLE

Chatting or typing speed is usually

Subgroup	Name, description
Word level	#Words (=W): number of words per minute; Avg. word length: average word length in a minute
Character level	#Chars (=C): number of characters in a minute
Punctuation	# ? and ! marks: * number of question and exclamation marks summed together, in a minute
Emoticons	Emoticons/C: number of emoticons divided by C # emoticons/W: number of emoticons divided by W.
Tempo/lexical	Char/Word writing speed: number of typed characters or words per second

calculated in terms of WPM. Words per minute, commonly abbreviated WPM, is a measure of words processed in a minute, often used as a measurement of typing speed or reading speed. For the purpose of typing measurement, each word is

standardized to be five characters or keystrokes long in English, including spaces and punctuation. For example, the phrase "I run" counts as one

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 12, March 2016

word, but "rhinoceros" and "let's talk" both count as two.

- **GROSS WPM**

Formula to find the Gross WPM

$$\text{Gross WPM} = \frac{\left(\frac{\text{All Typed Entries}}{5}\right)}{\text{Time (min)}}$$

Gross, or Raw WPM (Words Per Minute) is a calculation of exactly how fast you type with no error penalties. The gross typing speed is calculated by taking all words typed and dividing by the time it took to type the words in minutes.

- **NET WPM**

Formula to find the Net WPM

$$\text{Net WPM} = \text{Gross WPM} - \left(\frac{\text{Uncorrected Errors}}{\text{Time (min)}}\right)$$

$$= \frac{\left[\left(\frac{\text{All Typed Entries}}{5}\right) - \text{Uncorrected Errors}\right]}{\text{Time (min)}}$$

Net WPM is the most useful tool in gauging typing abilities. Since errors play a part in its calculation, it is more a measure of typing productivity than of just typing speed. In other words, a fast but error-prone typist will receive a lower net typing speed than a slower but more accurate typist - relatively speaking of course.

To calculate Net WPM, take your gross WPM result and subtract the amount of errors you left in per minute, also known as the error rate.

B. CHALLENGES IN CHATTING STYLE

Fig.2: Variations of chatting style.

a) Identifying Language

Chatting style may vary in the Languages most of the user may use their traditional language as their first language to chat with others. But it differs when chat with a different language peoples. Mostly everyone use the English as the main language to chat with others.

b) Usage of Shortcut

While chatting, a person may use the shortcut to chat with others the shortcut like LOL, AAF, B4, EOM, etc,. These are some of shortcut used by user to make a word easier to understand. This may challenge to count the words in the WPM. Mostly the person may be identified by the shortcut which personally had their personal style of chatting.

c) Variation of speed

User chatting style varies by the speed of typing the words. Chatting speed differs by the device using for chatting like PC, Mobile, Tab, etc., while using the different devices the chatting speed may vary. Measuring the speed of user typing as a WPM concept.

V. CONCLUSION

In this paper, we link up user accounts of the same natural person across different social network platforms. We propose a framework User chatting style, a multi-objective learning framework incorporating Heterogeneous behaviour core social network structure. We evaluate Chatting style against the state-of-the-art solutions on two real data sets - five popular Chinese social networks and two popular English social networks, a total of 10 million users and more than 10 terabytes of data. Experimental results demonstrate that USER CHATTING STYLE outperforms existing algorithms in identifying true user linkage across different platforms. In order to overcome the disadvantages of hydra we propose a new technique which measures the chatting styles of the persons and it provides a value for each word by comparing the values if it approximately matches then they are same user

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com
International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 3, Special Issue 12, March 2016

which they linked together, of various cross-platform.

REFERENCES

[1] Christo Ananth, M.Danya Priyadharshini, "A Secure Hash Message Authentication Code to avoid Certificate Revocation list Checking in Vehicular Adhoc networks", International Journal of Applied Engineering Research (IJAER), Volume 10, Special Issue 2, 2015,(1250-1254)

[2] J. Liu, F. Zhang, X. Song, Y.-I. Song, C.-Y. Lin, and H.-W. Hon, "What'sin a name?: an unsupervised approach to link users across communities,"in WSDM'13, 2013.

[3] R. Zafarani and H. Liu, "Connecting users across social media sites: Abehavioral-modeling approach," in KDD'13, 2013.

[4] S. Kumar, R. Zafarani, and H. Liu, "Understanding user migrationpatterns in social media," in AAAI'11, 2011, pp. -1-1.

[5] R. Zheng, J. Li, H. Chen, and Z. Huang, "A framework for authorshipidentification of online messages: Writing-style features and classification techniques," Journal of the Association for Information Science andTechnology, vol. 57, no. 3, 2006.

[6] T. Iofciu, P. Fankhauser, F. Abel, and K. Bischoff, "Identifying usersacross social tagging systems," in ICWSM'11, 2011, pp. -1-1.

[7] J. Zhang, X. Kong, and P. S. Yu, "Transferring heterogeneous linksacross location-based social networks," in WSDM'14, 2014, pp. 303-312.

[8] J. Wang, G. Li, J. X. Yu, and J. Feng, "Entity matching: How similaris similar," PVLDB, pp. 622-633, 2011

[9] Chats for all: A user survey to improve chats' interaction, Rocío Calvo, Ana Iglesias, Lourdes Moreno, Universidad Carlos III Computer Department Leganés, Spain

[10] Recognizing Chatting Style Rohan Puttagunta, Nick Wu, Renjie YouDecember 14, 2012

[11] just the way you chat:linking personality, style and recognizability in chats, Giorgio Roffo,Cinzia Giorgetta,Roberta Ferrario ,Marco Cristani,Universit` a degli Studi di Verona, Strada Le Grazie 15, I-37134 Verona, Italy.2ISTC-CNR, via alla Cascata 56/C, I-38123 Povo (Trento), Italy

[12] P. Jain and P. Kumaraguru, "@i to @me: An anatomy of username changing behavior on twitter," CoRR, 2014.

[13] A. Malhotra, L. C. Totti, W. M. Jr., P. Kumaraguru, and V. Almeida,"Studying user footprints in different online social networks," in ASONAM'12, 2012, pp. 1065-1070.

[14] A. Nunes, P. Calado, and B. Martins, "Resolving user identities over social networks through supervised learning and rich similarity features,"in SAC'12, 2012, pp. 728-729.

[15] J. Vosecky, D. Hong, and V. Shen, "User identification across multiple social networks," in NDT'09, 2009, pp. 360-365.

[16] N. Korula and S. Lattanzi, "An efficient reconciliation algorithm for social networks," PVLDB, pp. 377-388, 2014.

[17] X. Kong, J. Zhang, and P. S. Yu, "Inferring anchor links across multiple heterogeneous social networks," in CIKM'13, 2013, pp. 179-188.

[18] D. Koutra, H. Tong, and D. Lubensky, "Big-align: Fast bipartite graph alignment," in ICDM'13, 2013, pp. 389-398.

[19] G. Pickard, W. Pan, I. Rahwan, M. Cebrian, R. Crane, A. Madan, and A. Pentland, "Time-critical social mobilization," Science, vol. 334, no. 6055, pp. 509-512, 2011.

[20] R. T. Marler and J. S. Arora, "Survey of multi-objective optimization methods for engineering," Structural and multidisciplinary optimization, vol. 26, no. 6, pp. 369-395, 2004.

[21] R. Zafarani and H. Liu, "Connecting corresponding identities across communities," in ICWSM'09, 2009, pp. -1-1.