AN ANALYSIS OF USER IDENTIFICATION ACROSS MULTIPLE SOCIAL NETWORKS FOR SOCIAL IDENTITY RELATIONSHIP

ACHANA BALAMANI
Lecturer in Computer Applications,
Govt. Degree College, Begumpet,
Hyderabad. Email: balaa06@gmail.com

KINNERA VEERABAU
MCA, Dept. of Computer Science,
Saifabad PG College, Hyderabad.
Email: veeru.kinnera@gmail.com

ABSTRACT:

Social identity linkage from corner to corner diverse social media platforms is of critical prominence to business intelligence by acquisition from social data a deeper understanding and more accurate profiling of users. In this paper we suggests HYDRA framework with k-mean clustering which includes the social media networks which measures two users mention to one person when one of their attributes is same. The action of the user accounts are formed as a cluster by using k-mean clustering and thus the cluster has a data about the user where it mean to be efficient when proliferation of user increasing. Statistical models of topic distribution constructing structural consistency graph to evaluate the high-order structure consistency. Lastly, discovering the mapping function by multi-objective optimization compiled both the supervised learning and the cross platform structure consistency maximization. Henceforth, this model is able to find the hidden relationships of group of users with high delivery data speed.

Index Terms*: Social Identity Linkage, Structured Learning, optimization, structure consistency, user behavior trajectory.*

I. INTRODUCTION

The recent blossom of social network services of all kinds has revolutionized our social life by providing everyone with the ease and fun of sharing various information like never before (e.g., micro blogs, images, videos, reviews, location check-ins). 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 a deeper and better understanding of each individual user from the vast amount of Social data out there. 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 analysis is to link up all the data of the same user across different social platforms, offering the following benefits for user profiling. Completeness: Constrained by the features and design of each, any single social network 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. Continuity: While social platforms come and go, the underlying real-world users remain, who simply migrate to newer ones. User identity linkage makes it possible to integrate useful user information from those platforms that have over time become less popular or even abandoned. In this paper, we study the problem of automatically linking user accounts belonging to the same natural person across different social media platforms. It is beneficial to first explore the research challenges for a better understanding of this problem.

II. BACKGROUND WORKS

A. User Linkage across Social Media

User linkage was firstly formalized as connecting corresponding identities across

communities in and a web-search-based approach was pro- posed to address it. Previous research can be categorized into three types: user profile-based, user-generated-content-based, user-behavior model-based and social-structure-based. User-profile-based methods collect tagging information provided by users or user profiles from several social networks and then represent user profiles in vectors, of which each dimension corresponds to a profile field. Methods in this category suffer from huge effort of user tagging, different identifiable personal information types from site to site, and privacy of user profile. User-generated-content- based methods, on the other hand, collect personal identifiable information from public pages of user-generated content. Yet these methods still make the assumption of consistent usernames across social platforms, which is not the case in large-scale social networks platforms. User-behavior-model based methods analyze behavior patterns and build feature models from usernames, language and writing styles. Social- structure-based user linkage conduct linkage analysis by using structure features in social circles. For example, Korula et.al.solve the reconciliation of user's social network by starting from nodes with high degrees. Koutra et al. formulates the user linkage problem by learning an optimal permutation function between two graph affinity matrices. Based on user's social, spatial, temporal and text information, Kong et al. propose Multi-Network Anchoring to find the links between users from different platforms. Zhang et al. propose to predict heterogeneous links (social links and location links) inside the target social network given a set of anchor links among users from target network and source network. Previous methods

1) Seldom handle the missing information in usernames, user-generated content, behaviors and social structure; and

2) Have not given interpretation why there exists such missing information and how it impacts the user linking result.

B. Authorship Identification across Documents

Authorship identification is a task that identifies the authors of documents by their writing and language styles analyzed from their corresponding documents. Previous studies on authorship identification can be categorized into two types: content-based and behavior-model-based. Content-based-methods identify content features across a large number of documents. Behavior-model-based methods capture writing-style features, or build language models to identify content authorship. However, different from document scenario, social media platforms are much more complicated with multiple data media, graph/ social structures and missing information, which compromises most authorship identification methods.

C. Entity Resolution across Records

User linkage is in one way or another related to problems from other research com- munities including co-reference resolution in natural language processing, entity matching, graph node classification, record linkage in database, and name disambiguation in information retrieval, which can be generalized as entity resolution across records. Different from previous structure-based feature extraction approach and single feature based approaches, we consider a much more challenging setting where we examine multiple features a long time-line with missing and misaligned information and multiple media environments to link users across different platforms. Similarly, previous work on user identification on single site and deanonymization in social networks has been surveyed in, which are not elaborated here.

III. PROPOSED WORK

Design a new heterogeneous conduct model to degree the user conduct similarity from all factors of a consumer's social information. It is able to robustly deal with lacking data and misaligned conduct by using lengthy-term behavior distribution creation and a multi-resolution temporal conduct matching paradigm. The excessive heterogeneity of user social information can be preferred via the

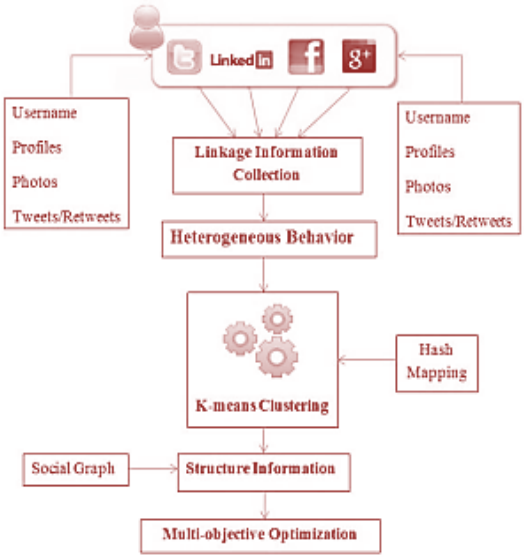
subsequent categorization of all the data approximately a user to be had.

User attributes: Included here are all the conventional based data about a user, e.g., demographic records, touch, and so on. The profile data is informative in distinguishing specific users. Common textual attributes in a user profile encompass name, gender, age, nationality, business enterprise, education, e-mail account, and so forth. A simple matching method may be constructed on such a fixed of information. However, the relative importance of these attributes are not same, because attributes consisting of gender and commonplace names like "John" aren't as discriminative as others consisting of e-mail cope with in identifying consumer linkage. The weights of the attributes used in the matching can be found out from large education set by probabilistic modeling.

User generated content material: Included right here are the unstructured information generated with the aid of users which includes textual content (reviews, micro-blogs, and so forth.), photographs, movies and so forth. Modeling is in most cases targeted at subject matter and style. A vital characteristic of social media platform is that over a sufficiently lengthy time period, the UGC of a user together offers a loyal mirrored image of the user's topical interest. Calculate the multi-scale temporal subject matter distribution inside a given temporal range for a user the use of the multi-scale temporal department. The intuition comes from the data that if two user talk over with one user, their tendencies have a tendency to be similar in the complete temporal range. Moreover, their tendencies in a shorter time period must also be similar. The more their dispositions are domestically matched in every shorter time period, the greater comparable their tendencies can be within the international variety. Thus the users are more likely to be the equal character.

User behavior trajectory: User conduct trajectory refers to all of the social conduct of a user as exhibited at the structures alongside the time-line, e.g., befriend, observe/unfollow, retweet, thumb-up/thumb-down, and so forth. The language style of a user together with personalized wording and emotion adoption is commonly properly meditated in feedback, tweets and re-tweets. To model user's function style, extract the most precise phrases of each person by means of a easy time period frequency evaluation at the complete database.

A. User core social networks features: A user's middle social networks are the social networks shaped among individuals who are the closet to the user, and the features are the aggregation of the user's middle social networks conduct. Social media sites with area-based-provider provide sturdy assist and incentive for recording and sharing user locations. On the other hand, comparable trajectory patterns across the platforms and no conflicting instances indicate the mobility similarity in actual international, as they would really like to offer test-in data on multiple social media structures. By reading the mobility similarity over a long length, a sufficiently excessive similarity in mobile trajectory means that the 2 users percentage comparable and even exactly the equal mobility behavior in real world. Therefore, the excessive mobility similarity may be considered as critical proof in social identity linkage.



System Architecture

B. Core Social Networks Features

Users tend to bring their closest friends over to different social platforms they frequently use. The behavior of a user's close friends is also informative in identifying different accounts of the same user. In the average similarity of the neighborhood data of two data items is more robust compared with the original similarity since it calculates the similarity of two convex hulls instead of two data points. Inspired by, this model the behavior of a user's social connections. Given two users i and i0 from different platforms, the behavior data of their top-k most frequently interacting friends are collected. If the similarity description between user i and i0, then a similarity vector is generated, including both the original similarity between i and i0, the average neighborhood similarity and the standard deviation of their social connection.

C. Structure Consistency Model

Propose a novel structure modeling method to maximize the behavior consistency on the users' core structure instead of user level behavior similarity. By propagating the linkage information along the social structure of each individual user, the model is capable of identifying user linkage even when ground-truth labeled linkage information is insufficient. 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, this model is able to consider the global structure between platforms to identify the true linkages and filter out those false ones. 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.

D. Multi-Objective Model Learning

Solve the social identity linkage (SIL) problem by multi-objective optimization (MOO) framework, where both the supervised learning on ground truth linkage information and the cross-platform structure consistency maximization are jointly performed towards Pareto optimality. Specifically, modify the formulations of kernel and linkage function, and develop a normalized-margin-based approach to deal with information missing in the similarity modeling. The linkage function by jointly minimizing the two objective functions via a unified multi-objective optimization framework. The model is a generalized semi-supervised learning approach by leveraging both ground truth linkage information and social structure.

HYDRA, a user linkage framework based on multi-objective optimization. It is composed of three main steps.

Step 1: Behavior similarity modeling. Calculate similarity among pairs of users via heterogeneous behavior modeling.

Step 2: Structure information modeling. Construct the structure consistency graph on user pairs by considering both the core network structure of the users and their behavior similarities.

Step 3: Multi-objective optimization with missing information. Construct multi-objective optimization which jointly optimizes the prediction accuracy on the labeled user pairs and structure consistency measurements across different platforms. The model is further modified to deal with significant information missing.

IV. CONCLUSION

Propose a framework, HYDRA, a multi-objective learning framework incorporating heterogeneous behavior model and core social networks structure. It performs even better than the baseline methods, and has better performance improvement with the increasing number of users. This shows that heterogeneous behavior model demonstrates better fitting to online social behaviors and social structure modeling helps to capture more linkable information. The distributed optimization method which optimizes the linkage functions in parallel on several servers

with a carefully designed synchronization strategy.

REFERENCES

[1] 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".

[2] R. Zafarani and H. Liu, "Connecting users across social media sites: A behavioral-modeling approach".

[3] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends R in Machine Learning, 3(1):1–122, 2011.

[4] J. Cai and M. Strube. End-to-end co reference resolution via hyper graph partitioning. In COLING'10.

[5] O. Hassanzadeh, K. Q. Pu, S. H. Yeganeh, R. J. Miller, M. Hernandez, L. Popa, and H. Ho, "Discovering linkage points over web data".

[6] I.Bhattacharya and L. Getoor, "Collective entity resolution in relational data".

[7] R. Zafarani and H. Liu, "Connecting corresponding identities across communities".

[8] S. Liu, S. Wang, and F. Zhu, "Structured Learning from Heterogeneous Behavior for Social Identity Linkage," IEEE Transaction Knowledge Data Engineering, vol. 27, no. 1, pp. 2005-2019, JULY. 2015.

[9] S. Liu, S. Wang, F. Zhu, J. Zhang, and R. Krishnan, "Hydra: Large scale social identity linkage via heterogeneous behavior modeling," in Proc. ACM SIGMOD Int. Conf. Manage. Data, pp. 51–62, 2014.

[10] T. Iofciu, P. Fankhauser, F. Abel, and K. Bischoff, "Identifying users across social tagging systems," in Proc. Int. Conf. Weblogs Social Media, p. 1, 2011.

[11] T. Iofciu, P. Fankhauser, F. Abel, and K. Bischoff, "Identifying users across social tagging systems".

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

[13] A. Malhotra, L. C. Totti, W. M. Jr., P. Kumaraguru, and V. Almeida, "Studying user footprints in different online social networks".

[14] A. Nunes, P. Calado, and B. Martins, "Resolving user identities over social networks through supervised learning and rich similarity features".

[15] J. Vosecky, D. Hong, and V. Shen, "User identification across multiple social networks".

[16] N. Korula and S. Lattanzi, "An efficient reconciliation algorithm for social networks".