International Journal of Computer Science and Engineering Communications,
Volume.4, Issue.2 (2016): Page.1378-1383
www.scientistlink.com/ijcsec

Analyzing the User Profile Linkage across Different Social Network Platforms

S.Monika[1], C.Anand[2], and R.K.Gnanamurthy[3]
PG Scholar[1], Assistant Professor[2], Professor[3]
Computer Science and Engineering[1,2]
K.S.R. College of Engineering, Tiruchencode, India.[1,2]
SKP Engineering College, Tiruvannamalai.India[3]
[1]moniinfotech93@gmail.com, [2]canand02@gmail.com

Abstract* — Social media analysis is to link up all the data of the same user across different social platforms, which is vital to business intelligence by gathering social data. This paper proposes HYDRA framework with k-mean clustering which comprises the social media networks which measures two users refer to one person when one of their attributes is identical. 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. Finally, discovering the mapping function by multi-objective optimization compiled both the supervised learning and the cross platform structure consistency maximization. Hence, this model is able to find the hidden relationships of group of users with high delivery data speed.

Keywords* — Heterogeneous behavior, k-mean clustering, multi-objective optimization, social identity linkage, structure consistency, user behavior trajectory.

I. INTRODUCTION

The rapid growth of the web in the last decade makes it the largest publicly accessible data source in the world. The amount of data/information on the web is huge and still growing. The coverage of the information is also very wide and diverse. One can find information on almost anything on the web. Data of all types exist on the web, e.g., structured tables, semistructured web pages, unstructured texts, and multimedia files (images, audios, and videos). Information on the web is heterogeneous. A significant amount of information on the web is linked. Hyperlinks exist among web pages within a site and across different sites. The web is dynamic. Information on the web changes constantly. Keeping up with the change and monitoring the change are important issues for many applications. Communication over social media networks, such as Facebook and Twitter, is gaining its importance in our daily life. Since the information exchanged over social media networks are not only texts but also URLs, images, and videos, they are challenging testbeds for the study of data mining. Probably the biggest and most challenging question concerning all businesses is how to leverage this big social data for better business intelligence. Unfortunately, information of a user from the present social network is fragmented, inconsistent and disruptive. The true power of social media is to link up all the data of the same user across different social platforms by offering the benefits of completeness, consistency and continuity to user profiling. A HYDRA framework with k-mean clustering is for cross-platform user identity linkage via heterogeneous behavior modeling. User behavior trajectory along temporal dimension: both empirical and social behavior studies over a sufficiently long period of time, a user’s social behavior exhibits a surprisingly high level of consistency across different platforms. A user’s core structures across different platforms share great similarity and offer a highly discriminative characterization of the user. Based on the behavior similarity model online users with multi-dimensional similarity vectors with the following information: a) the relative importance of the user attributes, which measures how likely two user accounts belong to the same person when any one of their attributes is identical; b) the action of the user accounts are formed as a cluster by using k-mean clustering and thus it to be efficient when proliferation of user increasing; c) the statistical divergence of topic distributions, describing the potential inclination of users over a long period of time; d) the overall degree of matching between users' behavior trajectories, capturing the highly correlated actions between user accounts over a certain period of time.

II. RELATED WORKS

A. Studying User Footprints

In the digital age, when all are living two lives, offline and online, paper trails and digital trails coexist. Digital trail captures the interactions and behaviors in the digital environment. The online digital footprints capture the online identity and whatever to do on the web becomes part of our online identity forever. By uniquely identifying users across social networks can discover and link her multiple online profiles. Linking together users multiple online identities has many benefits e.g. profile management like managing setting and building a global social networking profile, help user monitor and control the personal information leakage, user profile portability, personalization. User's digital footprints within a service are the set of all information related to user, which was either provided by the user directly or extracted by observing the user's interaction with the service.

B. Inferring Anchor Links

Different from link prediction methods, the approach, called MNA (Multi-Network Anchoring), can extract heterogeneous features from multiple heterogeneous networks for anchor link prediction, including user's social, spatial, temporal and text information. This extended some existing social features for link prediction into multinetwork settings, based upon the known anchor links. Users often have similar social links in different social networks, such as Twitter and Facebook, because such social links usually indicate the user's social ties in real life. That can make use of the social similarity between two user accounts from different social networks to help locate the same user.

C. Connecting Users across Social Media Sites

A behavioral-modeling approach introduces a methodology (MOBIUS) for finding the mapping among identities across social media sites. The methodology is based on behavioral patterns that users exhibit in social media, and has roots in behavioral theories in sociology and psychology. Unique behaviors due to environment, personality, or even human limitations can create redundant information across social media sites. The methodology exploits such redundancies to identify users across social media sites. It only uses the minimum amount of information available across sites. MOBIUS contains 1) behavioral patterns, 2) features constructed to capture information redundancies due to these patterns, and 3) a learning framework.

D. Discovering Linkage Points over Web Data

Discovering Linkage Points is a framework for identification of linkage points for multi-source Web data integration. The framework includes a novel class of search algorithms to identify strong linkage points even when such attributes are weak schema matches. Importantly, they are specifically looking for attributes that help in identification of entities that can be linked. So unlike in schema matching, they are not interested in finding all corresponding attributes (for example, matching color and colour). As a result, the search can be much more focused. The framework for linkage point discovery takes input data sets and returns a (ranked) set of linkage points.

E. Predicting Social Links across Aligned Heterogeneous Social Networks

The link prediction problem for new users is different from traditional link prediction problems. Conventional supervised link prediction methods implicitly or explicitly assume that the information are identically distributed over all the nodes in the network without considering the joining time of the users. The models trained over one part of the network can be directly used to predict links in other parts of the network. As a result, conventional supervised link prediction models trained over old users based upon structural features. Predicting social links among accounts of the same user as "anchor links", which could help align user accounts across multiple social networks. New users in one social network might have been using other social networks for a long time. These user accounts in the source networks can provide additional information about the new users in the source network. The target network and the source network, with aligned user accounts. The target network, there are many old users with abundant social links and auxiliary information, such as posts, spatial and temporal activities. Exploit the new user’s information in source networks to help improve the link prediction performances in the target network. The problem of social link prediction for new users by using aligned social networks is novel and totally different from existing link prediction problems, e.g., link prediction via personalized sampling, pseudo cold start link prediction problem, link prediction via network transfer and the traditional transfer learning problems in feature space. Social link prediction for new users across aligned social networks is very challenging to solve due to the following reasons:

- **Differences in information distributions:** In order to use the old user’s information in the target network, overcome the problem of the differences in information distributions between old users and new users.
- **No auxiliary information:** Another key part of the problem study is the cold start link prediction problem caused by the lack of information about these new users.
- **Aligned social networks:** Works on transfer learning focus on transferring knowledge between two domains via shared feature space or between two networks through shared triad linkage structures.

III. PROPOSED SYSTEM

A. Heterogeneous Behavioral Model

Design a new heterogeneous behavior model to measure the user behavior similarity from all aspects of a user’s social data. It is able to robustly deal with missing information and misaligned behavior by long-term behavior distribution construction and a multi-resolution temporal behavior matching paradigm. The high heterogeneity of user social data can be appreciated by the following categorization of all the data about a user available.

User attributes: Included here are all the traditional structured data about a user, e.g., demographic information, contact, etc. The profile information is informative in distinguishing different users. Common textual attributes in a user profile include name, gender, age, nationality, company, education, email account, etc. A simple matching strategy can be built on such a set of information. However, the relative importance of these attributes are not identical, because attributes such as gender and common names like “John” are not as discriminative as others such as email address in identifying user linkage. The weights of the attributes used in the matching can be learned from large training set by probabilistic modeling.

User generated content: Included here are the unstructured data generated by users such as text (reviews, micro-blogs, etc.), images, videos and so on. Modeling is primarily targeted at topic and style. An important feature of social media platform is that over a sufficiently long period of time, the UGC of a user collectively gives a faithful reflection of the user’s topical interest. Calculate the multi-scale temporal topic distribution within a given temporal range for a user using the multi-scale temporal division. The intuition comes from the fact that if two users refer to one person, their inclinations tend to be similar in the whole temporal range. Moreover, their inclinations in a shorter time period should also be similar. The more their inclinations are locally matched in every shorter time period, the more similar their inclinations will be in the global range. Thus the users are more likely to be the same person.

User behavior trajectory: User behavior trajectory refers to all the social behavior of a user as exhibited on the platforms along the time-line, e.g., befriend, follow/unfollow, retweet, thumb-up/thumb-down, etc. The language style of a user including personalized wording and emotion adoption is usually well reflected in comments, tweets and re-tweets. To model a user’s characteristic style, extract the most unique words of each user by a simple term frequency analysis on the whole database.

User core social networks features: A user’s core social networks are the social networks formed among those who are the closet to the user, and the features are the aggregation of the user’s core social networks behavior. Social media sites with location-based-service provide strong support and incentive for recording and sharing user locations. On the other hand, similar trajectory patterns across the platforms and no conflicting instances indicate the mobility similarity in real world, as they would like to provide check-in information on multiple social media platforms. By analyzing the mobility similarity over a long period, a sufficiently high similarity in mobile trajectory implies that the two users share similar and even exactly the same mobility behavior in real world. Therefore, the high mobility similarity can be considered as important evidence in social identity linkage.

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 i_0 from different platforms, the behavior data of their top-k most frequently interacting friends are collected. If the similarity description between user i and i_0, then a similarity vector is generated, including both the original similarity between i and i_0, the average neighborhood similarity and the standard deviation of their social connection.

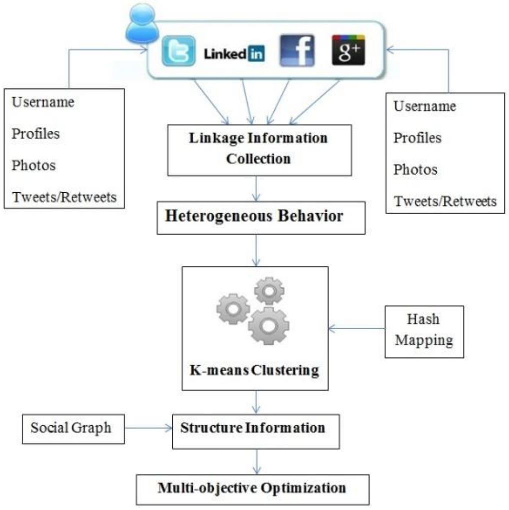


Fig. 1. System Architecture

The average similarity features and the standard deviation features measure the inclination and the behavior consistency of the friend groups.

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.

IV. FRAMEWORK OVERVIEW

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

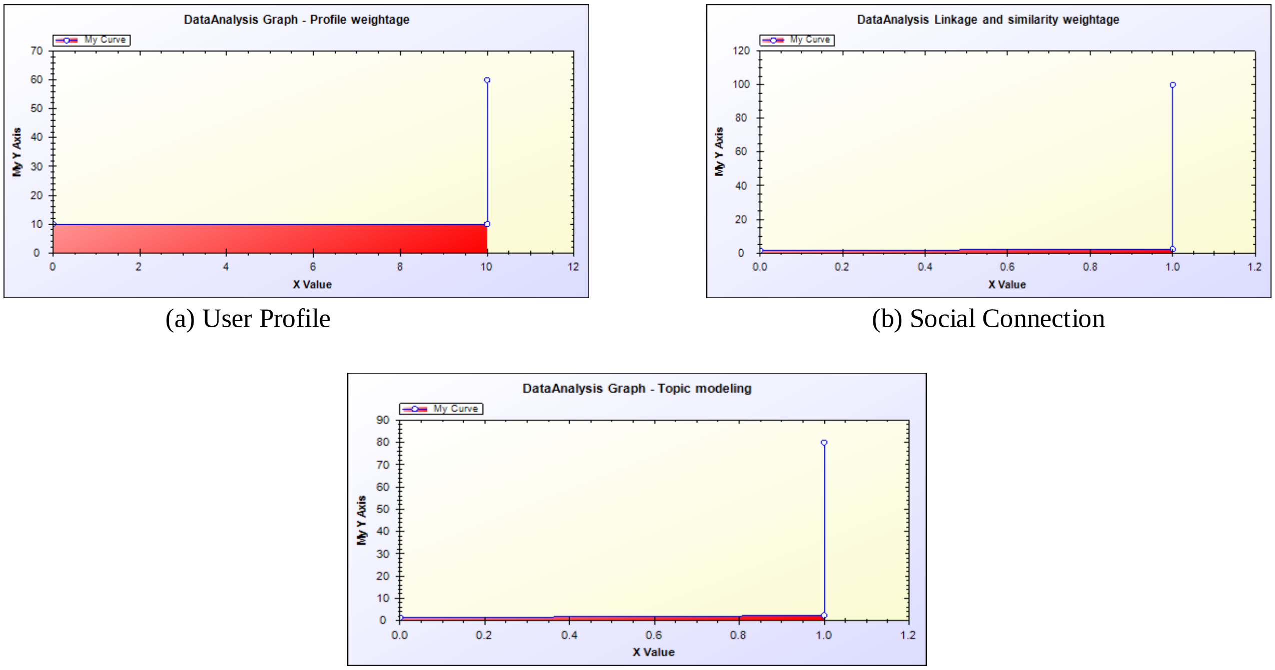
Step 1: Behavior similarity modeling. Calculate similarity among pairs of users via heterogeneous behavior modeling.

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.

V. EXPERIMENTAL RESULT

The social data sets used for the experiments which include four different platform services. Users in this data set each with accounts on multiple social sites like facebook, twitter, etc., In that social networks, collect user profiles (e.g. gender, city, and favorites), social content (e.g. tweets, posts, and status), social connections (e.g., friendship, comments, and repost or retweet contents), and timeline information (e.g., time index for each behavior) to optimize the information matching. The 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 benefits to user profiling. Single social networks service offers only a partial view of a user from a particular perspective is referred as completeness. Consistency provided information by users on a social platform could be false, conflicting, missing and deceptive. Continuity which refers the social platforms comes and go, the underlying real persons remain, and simply migrate to newer ones.



(a) User Profile

(b) Social Connection

(c) Timeline Information

Fig. 2. Analyzing the Weightage of User Profile, Identity Linkage and Topic Modeling

The ground truth of the linkage of each user across all the platforms is provided to all accounts on different websites. The most reliable data is to uniquely identify a natural person and link all the different accounts. The input design must ensure user-friendly screens, with simplicity, providing ease of viewing & entering the data of the main objective designing.

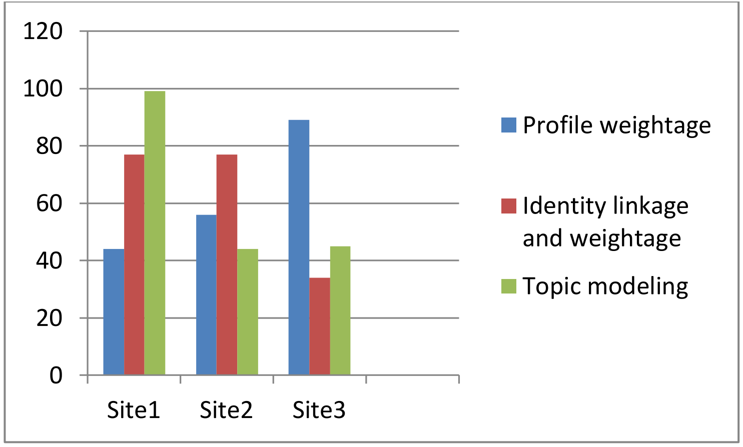


Fig. 3. Dimensional Graph of Individual Sites

Compare all the sites with their profile, identity linkage and topic modeling which are to achieve better performance of all the approaches, a validation set with four platform user pairs and their ground truth labels with k-means clustering have been used. The output result is to provide an environment that is both convenient and efficient to use in retrieving and storing data information.

VI. CONCLUSION

In this paper, the information of a user from the current social scene is fragmented, inconsistent and disruptive so link up user accounts of the same natural person across different social network platforms. 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]A. Malhotra, L. C. Totti, W. M. Jr., P. Kumaraguru, and V. Almeida, “Studying user footprints in different online social networks,” in Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Analysis and. Mining, pp. 1065–1070, 2012.

[2]A. Nunes, P. Calado, and B. Martins, “Resolving user identities over social networks through supervised learning and rich similarity features,” in Proc. 27th Annu. ACM Symp. Appl. Comput., pp. 728–729, 2012.

[3]J. Zhang, X. Kong, and P. S. Yu, “Transferring heterogeneous links across location-based social networks,” in Proc. 7th ACM Int. Conf. Web Search Data Mining, pp. 303–312, 2014.

[4]O. Hassanzadeh, K. Q. Pu, S. H. Yeganeh, R. J. Miller, M. Hernandez, L. Popa, and H. Ho, “Discovering linkage points over web data,” Proc. VLDB Endowment, vol. 6, no. 6, pp. 444–456, 2013.

[5]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, pp. 41–49, 2013.

[6]S. Kumar, R. Zafarani, and H. Liu, “Understanding user migration patterns in social media,” in AAAI’11, pp. –1–1, 2011.

[7]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.

[8]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.

[9]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.

[10]W. Cui, Y. Xiao, H. Wang, Y. Lu, and W. Wang, “Online search of overlapping communities,” in Proc. SIGMOD Conf., pp. 277–288, 2013.