ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

Similarity based Matching Approach for User Identity across Multiple Social Platforms

Divya D S[1], Dr. Asha. T[2]

Student, Department of Computer Science, Bangalore Institute of Technology, Bangalore, Karnataka, India [1]

Professor, Department of Computer Science, Bangalore Institute of Technology, Bangalore, Karnataka, India[2]

ABSTRACT: Now a days linking the social identity across the different social media platforms is of critical importance to business intelligence. In this paper, we propose a solution framework called Similarity based matching approach, which consists of three key steps. First we model heterogeneous behavior by long-term topical distribution analysis and multi-resolution temporal behavior matching against high noise and information missing. The behavior similarity are described by multi-dimensional similarity vector for each user pair. Second we build structure consistency models to maximize the structure and behavior consistency on user's 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. Third we propose a normalized margin based linkage function formulation, and learn the linkage function by multi-objective optimization. Here both the supervised pair-wise linkage function learning and structure consistency maximization are conducted towards a unified Pareto optimal solution. The proposed methodology is able to deal with user information across the multiple platforms, and correctly identifies the real user linkage from noisy user behavior data records.

KEYWORDS: Social identity linkage; behavior similarity; parento optimization; supervised function; temporal matching behavior.

I. INTRODUCTION

People have accounts in different social network platforms like Face book, Twitter, and Orkut etc. They maintain their account under different identities. It's very difficult to find the single user identity among these network platforms. Since, in cyber network also nobody has noticed these fraudulent activities. Now a days, everyone shares their information through social network services. At the same time, the biggest question arises is how to get these users big data for identifying the actual user. So people were thinking how to get those individual user information from multiple platforms.

But due to same person maintaining accounts in different social media platform, using the logs of one platform alone to get the user behavior is not a good practice. So there should be a way to connect the different social media accounts to single user and collect all the records for that user from all platforms. Only then correct user behavior and interests has been inferred. In this project, we provide a solution for linking different accounts in different social media platforms to a single user. This can provide the various benefits like cross checking the user identity would enables us to thorough understanding of a each individual user's information and the structure of their behavior patterns.

II. RELATED WORK

Reza Zafarani et al [1] studied the user linkage across the various communities and based on the web search approach it turns out that usernames across the different communities can be identified and the corresponding usernames in various communities have been studied. T. Iofciu et al [2] investigate whether users can be identified across social tagging systems and collect the user tagging information for detecting the actual user and these user profiles can be indicated in vector form. N.Korula et al [3] authors designed the local distribution algorithm for social network reconcillation to link the large set of user identification nodes in two different network graphs. For this purpose only the structural details can be used. R. Zheng et al [4] proposed a framework for identifying the

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

authorization of the user and messages sent by the user through online. This framework can adopt the punctuations used in the text and semantic features etc and it also combine the functions used in the words. A. Nunes et al[5] describes an approach for solving the user identity problems by comparing the various features and attributes of the user. This can classify the user data in binary form to find whether the data belongs to same.By combining these user or not feature vectors, the similarity score is calculated. For example, similarity between user basic information such as name, gender, age, date of birth and user visited location etc. J. Liu et al [6] proposed semi supervised learning framework to map common users across social networks and the same user profiles in the different platforms to check for the similarity has been inferred. For each network the graph can be drawn and later it build the semi supervised framework to compare between the two users. S. Liu et al [7] proposed HYDRA, this is a framework for cross checking the user accounts to identify the real user by using the user behavior and their interests. This framework can take the significance of two unique features such as user trajectory modeling and core structures of the user that has been posted by the user in social networks. J. Vosecky et al [8] proposed the user identification methods to find the user's personal details for a user's personal details on Facebook, twitter and describes a group of methods to exploits required information of user on multiple platforms, to search for each individual user identity on Facebook. The methods include user basic data, trajectory information and social network features.

III. PROPOSED METHODOLOGY

In proposed system we design a similarity based matching approach, a framework for cross checking the users information among the veracious platforms that makes us to find the real user by using HYDRA architecture. The input to the system is user data across the social media and the output of the system is, the real user identified across the multiple networks.

The user data attributes is divided into five modules.

1) Topic Distribution modeling
2) User style modeling
3) User Trajectory modeling
4) User behavior modeling
5) Social features modeling

1. Topic Distribution Modeling

This module includes the unstructured data generated by the user such as text, images, etc. In this User Topics modeling we collect all the messages sent by the user and those messages can be grouped into the categories by using Naïve Bayes classifier.

2. User Style Modeling

In this module, each word in the messages that has been exchanged between the users and the emotions he/she has used can be helpful to distinguish between the users. Here we collect only the most specific words by removing all the stop words such as a, an, the etc. After collecting the most specific words we can calculate the similarity score by dividing the matched words by the total number of words i.e. k.

Then we calculate a similarity score S, for user pairs by using the following formulae

$$S = \text{number of matched words} / k$$

3. User Trajectory Modeling

This is one of the important module in finding the similarity matching between the users. The users on the different social platforms can post same or duplicate information over the social accounts.

Hence, this user trajectory can be considered as an one of the important evidence in finding the similarity.

4. User Behavior Modeling

In User Behavior Modeling, the similarity score is calculated based on the user friends that are communicated frequently with each other. Here we can collect the close friends of the user from different accounts. This information is very helpful in finding the actual user. The similarity score of two users on different accounts is more robust in calculating the similarity vector between the users.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

5. Social Features Modeling

The Social Features module is the last and one of the important module in finding the similarity between the users. Here the similarity is calculated based on how many messages has been exchanged between the users. Then we calculate the difference between those messages and that difference is matching with the fixed threshold value then we say that two users belongs to the same person.

Similarity Matching

Once the similarity scores of all the modules have been completed, we have to check for whether the two users belong to the same person. For that we have fixed one threshold value, if the score of each module is greater than that threshold then we find out the real user.

IV. SIMULATION RESULTS

This model contains the user data from two social platforms like Face book and twitter. Here we collected the one user's data from the two multiple accounts such as Twitter and the Face book, similarly the another user's data from the same two accounts. We divided the user data into five different modules based on the user attributes. By comparing each of the user attribute at different accounts we can find out the real user.

The Fig. 1 shows the initial GUI where the user text document can be loaded. Here the extracted user data can be uploaded from collected data files. The Fig. 2 shows the User1 profile data from Facebook, in which we divided the user profile information into five different attributes. The Fig. 3 shows the User1 profile data from twitter, in which we divided the user profile information into five different attributes. The Fig. 4 shows the vectorization of the Topic distribution modeling and User style modeling. In topic distribution modelling, all the messages sent by the user are put into the particular group based on naïve bayes classifier. In user style modeling, all the specific words can be extracted from user messages by eliminating the stop words. The Fig. 5 illustrates the similarity score of each module. The score vector for the above two user profiles has been calculated and displayed here. We fixed one threshold value i.e 0.5, and the similarity score of all the attributes should be greater than this value. Here the calculated similarity score is 0.8. Hence this value is greater than threshold value and we concluded that two users belongs to same person. The Fig. 6 demonstrates the performance analysis graph. It illustrates the time taken by the each user module to perform the similarity score. The x-axis indicates the user attributes and Y-axis indicates time taken by the each attribute.

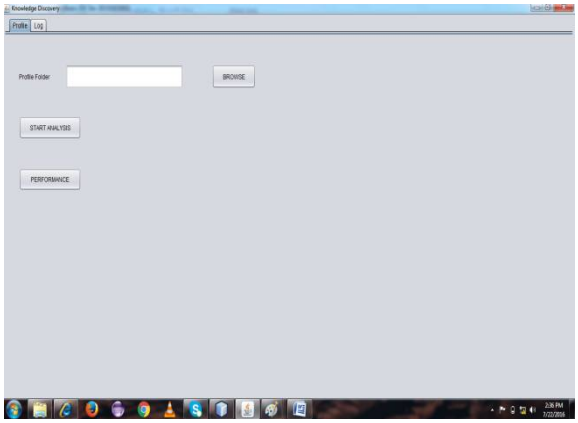

Fig. 1: Initial GUI to load the user text document

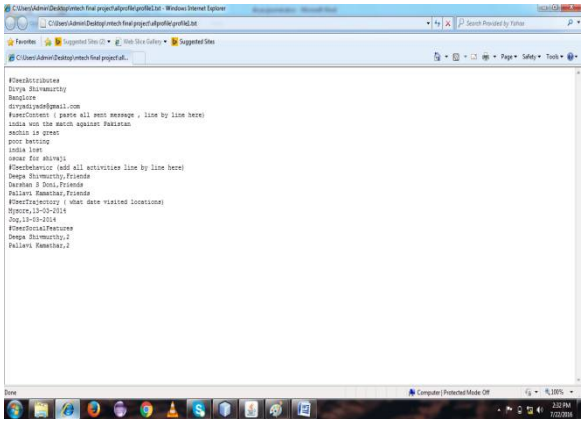

Fig. 2: User1 Profile Data

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

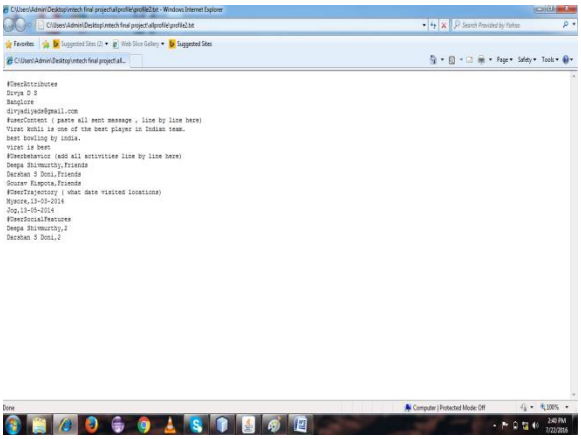

Fig. 3: User2 Profile Data

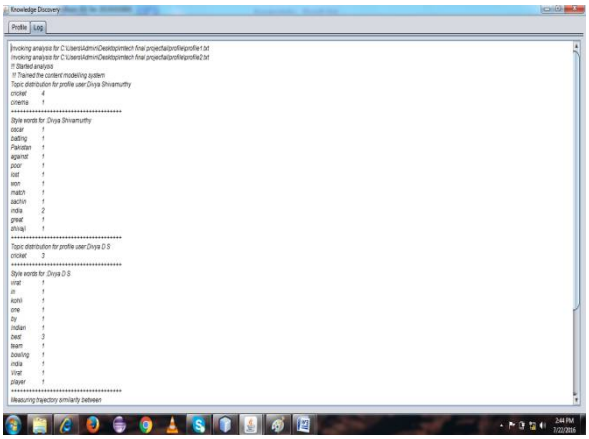

Fig. 4: Vectorization of user profile Data

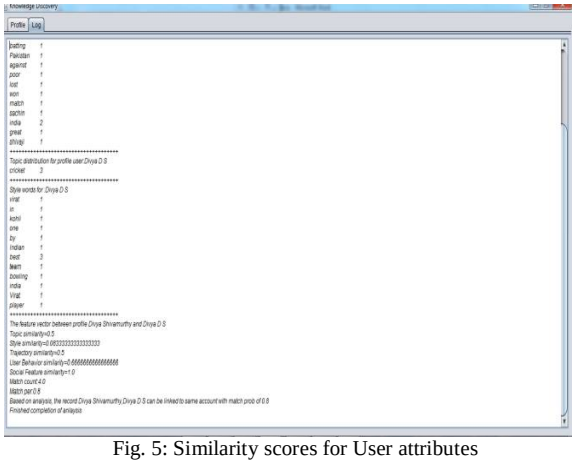

Fig. 5: Similarity scores for User attributes

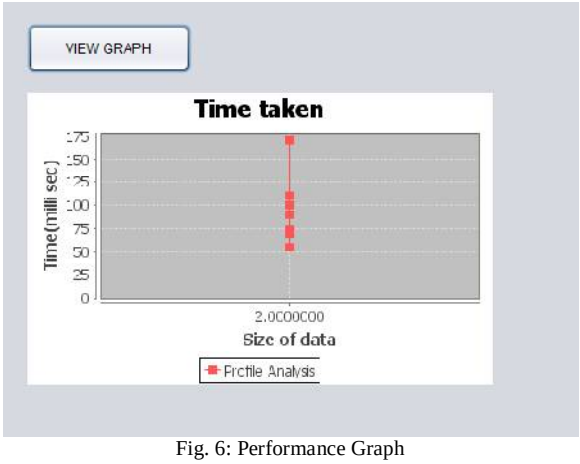

Fig. 6: Performance Graph

V. CONCLUSION AND FUTURE WORK

In this paper, we proposed a framework called Similarity based matching approach using HYDRA architecture, to find a real user among two different social networks such as Face book and twitter. The proposed methodology is very efficient in identifying the real user by comparing all the attributes of a user among multiple social networks.

In experimental analysis we conduct the Similarity based matching approach for two different data sets and calculated the similarity score of each attribute. By comparing these similarity scores with fixed threshold value we conclude that two users belongs to same person.

REFERENCES

[1] Reza Zafarani and Huan Liu., "Connecting users across social media sites: A behavioral-modeling approach," international conference on Knowledge discovery and data mining, pp.41-49, 2013.
[2] T. Iofciu, P. Fankhauser, F. Abel, and K. Bischoff, "Identifying users across social tagging systems," international conference on weblogs and social media, Barcelona, Spain, 17-21 july 2011.
[3] N. Korula and S. Lattanzi, "An efficient reconciliation algorithm for social networks," PVLDB, Volume 7, Issues 5, pp.377–388, 2014.
[4] R. Zheng, J. Li, H. Chen, and Z. Huang, "A framework for authorship identification of online messages: Writing-style features and classification techniques," Journal of the Association for Information Science and Technology, vol. 57, Issues 3, pp.378-393, 2006.
[5] A. Nunes, P. Calado, and B. Martins, "Resolving user identities over social networks through supervised learning and rich similarity features," in proceedings of the 27th Annual ACM Symposium on Appied computing. pp.728-729, 2012.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

[6] 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 proceedings of the sixth ACM international Conference on Web search and Data Mining , pp. 495-504, 2014.
[7] S. Liu, S. Wang, F. Zhu, J. Zhang, and R. Krishnan, "Hydra: Large-scale social identity linkage via heterogeneous behavior modeling." in proceeding of the 41st ACM SIGMOD International Conference on Management of Data, pp.51-62, June 22-27, 2014.
[8] J. Vosecky, D. Hong, and V. Shen, "User identification across multiple social networks," ACM Transactions on knowledge Discovery from Data(TKDD), volume 10, Issue 2, October 2015.

BIOGRAPHY

Divya D S is currently pursuing M.Tech degree from Bangalore Institute of Technology, Bangalore, Karnataka, India. She has obtained her B.E degree from Sri Siddhartha Institute of Technology and Management, Tumkur, Karnataka. Her research interests are Big Data and data Mining.

Dr. Asha.T obtained her Bachelors and Masters in Engineering, from Bangalore University, Karnataka, India. She has her Ph.D from Visvesvaraya Technological University under the guidance of Dr. S. Natarajan and Dr. K.N.B. Murthy. She has over 20 years of teaching experience and currently working as Professor & PG Coordinator in the Dept. of Computer Science & Engg., B.I.T. Karnataka, India. Her research interests are Data Mining, Medical Applications, Pattern Recognition and Artificial Intelligence.