Identifying Linkage Across Social Network Platforms based on Heterogeneous Modeling

Sowmya A
8th Sem, ISE, EPCET
Bangalore

Anitha N
Assoc Prof, Dept of ISE
EPCET, Bangalore

***Abstract*—People pose with different identities on various social media, this is of critical importance for most of the business that target social media to gain popularity as well as for the cyber security. Here we propose a framework, HYDRA, which has the following ways to provide a solution (I) we use the long time behavior of the profile of the users and match the information missing along with the behavior of the user (II) Since user's have account in various platforms we have to get the accurate information so the model should be built for consistency (III) We learn the profiles using the function formulation and linkage to get the parento similarity. Here we design a model that can handle various problems such as consistency, information missing along with the real-linkage across the different platforms**

I. INTRODUCTION

The cyber virtual world has been giving importance to see whether people use their true identity across various online social networks, and also monitor if they pose any other identity that is the critical importance as to have a clear route so that they don't pose any harm to other people across the social media. So people feel safe in sharing the videos, photos and other information with others. Here a lot of data is generated across different platforms that have to be maintained, which at times are inconsistent, disruptive and also fragmented. The main aim of the proposed project is to get the benefits of user profiling, i.e., linkage of the user's data and information across platforms along with the following: **Consistency—** sometimes the information provided by the user's [1] may be false, deceptive, missing or conflicting such problems can be solved by cross- checking the user's identity across various different platforms.

Missing Information—most of the user's profile attributes are incomplete, i.e. they have a very few user whose profile attributes are complete by providing all the information required. Such information missing could lead to inconsistency and also real-linkage finding across cross-platforms.

Completeness—through linking various user profiles across platforms we get the complete information of the user compared to very little information from one single profile. **Continuity--**people move from one social platform to another as time advances leading to abandonment of the older social platforms and migrating to newer ones. So by linking the user's profile the identity remains the same across various cross-platforms:

Data Misalignment—Users may be comfortable in using a few of the platforms regularly and a few times. For ex. A few of them will post the images "family and friends" on the Face book and that of their "professional life" on other social media like LinkedIn. People use various media like the images, blogs, tweets, status update etc. Data posted by the user's many have a certain privacy policy activated like private for photos and public for the status update and also timely updates might be varying across the different social platforms leading to inconsistency of data present in the current.

Undependable Attributes— since the social platforms is being accessed globally the way people represent themselves is quite unique like for ex: their names and surnames a few prefer last name in the beginning followed by their first name afterwards [2], which creates some confusion in obtaining the name of the user. Some undependable attributes are the age and gender, where most of the time the given attributes are false. This leads to inaccuracy in obtaining the information for linkage and the information obtained by structured learning.

In the proposed framework HYDRA, we use the social data for obtaining the User's behaviour trajectory and also the user's core social platform features. The study of the online social platform over a period of time shows that the user's behaviour is similar across the different platforms and also the core features of the social network which is being frequently used.

To obtain the similarity between the user profiles, we use the following details such as the user's attributes that is primarily in matching the user profile, the content generated by the user over a period of time and the behaviour trajectory of the user. We create a linkage among the users based on the information provided by them and matching them using the social structure by function formulation methodology.

II. RELATED WORK

In this paper [3] , the author made a investigation on the questions regarding the identification of users across the social systems by combining two informations such as user ids and their respective tag data. They also introduced different approaches and compared them with respect to measurement of distance between different user profiles. They were able to achieve around 60% - 80% accuracy depending on the settings based on the best combination. Hence they concluded that the web 2.0 traces of users could able to reveal most of the user identity.

An identity of a user on an online social network (OSN) is defined [4] by their profile content and network attributes. OSNs allow users to change their online attributes with time to reflect changes in their real-life. Temporal changes in user's content and network attributes have been well studied in literature, however little research has explored temporal changes in profile attributes of online users. This work makes the first attempt to study changes to a unique profile attribute

of a user – username and on a popular OSN which allows users to change usernames multiple times – Twitter. We collect monitor and analyze 8.7 million Twitter users at macroscopic level and 10,000 users at microscopic level to understand username changing behaviour.

In this paper [5], to identify the user from one social network with another, they applied some automation techniques on the online footprints which were in the digital form of the user's. Then they extracted this footprint completely from the public profile of the user information. Later, using different similarity metrics, they compared different profile features and identified the discrimination and also assessed the discriminative capacity. They compared Name and UserID with the Jaro Winkler metric, which found to be the best discriminative method. So they could achieve precision of 99%, accuracy of 98% and recall of 96% by using the best similarity metrics and the best set of features.They also performed testing on real world data of the system to check the user profiles using the name displayed on LinkedIn user of Twitter. This testing resulted in 75% time the display of accurate answer which was the top three correct profiles. Hence this proposed user profile discriminative system could help for security analysts and also to analyse and compare two social networks. Their future plan is to include three to four profile fields and to generalize the model so that it can be applicable to other social networks.

In recent years, Location-based social networks (LBSNs) are being attracted and gaining much attention as it offers geographic services It has heterogeneous nodes and many links, hence they are very complex structures. The prediction task of location links and social links in real world LSBN are strongly influential and strongly correlated. Whenever LSBN branches to new social groups or new geographic areas, it encounters data sparsity problem in link prediction, which is one of the key challenge.

In this paper [6], the author proposes a new method called TRAIL (Transfer heterogeneous links across LBSNs), to solve the problem of simultaneously predicting multiple types of links for a new LBSN across partially aligned LBSNs. For new posts on location links and social links, this new technique- TRAIL can extract heterogeneous features and also from online posts it can gather information. Also simultaneously it has the capability to predict multiple types of links. To solve the problem of lacking information, it can also transfer information from one network to the other new network. Through experiments performed on two real-world LBSNs proved that TRAIL achieves very good performance.

In this paper [7], to merge human judgement and machine learning, they proposed a new method called LOAD(Labelling Oriented Author Disambiguation). Its framework consists of both high recall and precision clustering, and also top dissimilar clusters selection with ranking. To train the similarity functions between publications supervised learning algorithms are incorporated in the framework. Further to generate clusters, clustering algorithms are also used. Comprehensive experiments were performed on LOAD to validate the efficiency and effectiveness. Results revealed that LOAD yields more accurate results when ccompared to conventional algorithms for assisting human labelling.

The author in this paper [8] has first provided the definition of clustering to introduce to the basic concept of clustering. Then they have provided different approaches to data clustering along with algorithms to implement those approaches. The major issue is how to retrieve relevant information quickly from the databases. Data clustering is one of the different techniques that have been developed for this purpose. Hence the process of clustering is to filter results obtained from search engine and provide accurate results. Hence this paper concentrates more on Data Clustering and the different approaches and along with their analysis.

In present arena, more people have their virtual identities on the web [9]. It is noticed that more people are users of more than one social network and even their friends may be registered on multiple websites. To enable user with up-to-date information such as virtual contacts, a procedure is needed to aggregate the online friends into a single integrated environment. Also an improved procedure is needed to search for people across different websites. In this paper, they proposed a approach for identifying the users based on profile matching. To study the similarity of different profile definition, they used data from two popular social networks. They found out the importance of different fields in the web profile. They developed a profile comparison tool and with this tool, they could identify and consolidating duplicated users on different websites, which proved the effectiveness and efficiency of the developed tool.

III. PROPOSED WORK

The heterogeneous model built here can deal with the missing information and also the behaviour of the user's on the social network that is observed for learning over a long period of time. The user's behaviour in the network is carefully watched for the consistency so that the structured learning is obtained for maximum potential. The linkage across the social platform for the learning the user's across cross-platform, consistent data all these work for the parento similarity dealing with the problems that is caused during the linkage.

HYDRA has been experimented using the real big data sets over the social cross-platform globally spread network that has led to satisfactory results than the previous method used that were platform dependent.

Identity Linkage for the user Profile:- let us consider a set of people say M who has accounts on the social platforms T and T'. Now the task is to check whether M1 and M2 profile reflects the same person or not.

$$f(u1,u2) = \begin{cases} 1, & \text{if } M1=M2 \\ 0, & \text{otherwise} \end{cases}$$

This mathematical formulation will lead to the high computational cost since pair wise computation is required, which grows exponentially with the large set of user and the cross-platforms. But it produces sufficient data to be matched and also the consistency is being maintained for the user's behavior in the social group. HYDRA provides a platform that comprises the behavior modeling for the parento similarity, structured modeling through learning the user's profile for a long period of time and also the missing information that has been filtered through the linkage across various social platforms.

A. Heterogeneous Behavior Model

The user's generates a large amount of data (images, videos, text etc) all these are heterogeneous in nature across the various social networks. These lead us to greater missing

of information because of core social network features. The HYDRA architecture is shown in figure 1 and explained as follows:

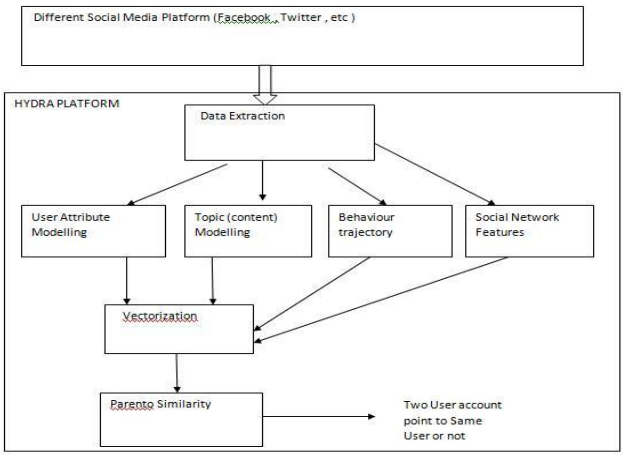

Fig 1: The HYDRA architecture

- **User Attributes: -** the attributes gives us the general information about the user ex. Name, email, gender, nationality etc... These are been matched with other accounts for the similarity check. However these are not sufficient since people have common names across the globe that uses the same name as the user name to login. Along with these the visual attributes are also taken into account since the profile can also be matched using the face recognition technique.
- **User Generated Content:-** These are the information that is been obtain from the user's generated contents like the tweets, images, topic of interest over a period of time dividing them to be the interests with some particular space.
- **User Style Modeling: -** The language usage of each user is also monitored with some style words. Later these words are been checked for their usage by removing the stop words (i.e. the, an, it etc).
- **User Behavior Trajectory: -** This is to check for the various interest of the user's like pages that are been liked/unlike, adding friends or un friending, follows etc.
- **User Core Social Network Feature:-** Most of the people who interact in a social media get along with those who are present within the group this gives us the information or idea about the user's profile . We use the similarity dimension for these so as to compare between the accounts.

IV. EXPERIMENTAL RESULTS

The front end works along with the back end for the extraction of the data from the user profile. The following snapshots represent the output of the program modules of the system. The figure 2 depicts the analysis of all the profiles collected

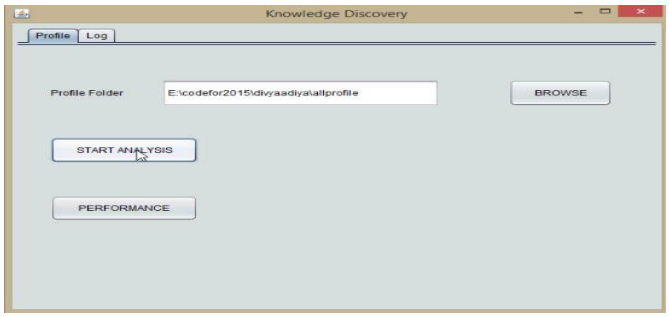

Fig 2. Represents the analysis stage for the given profile

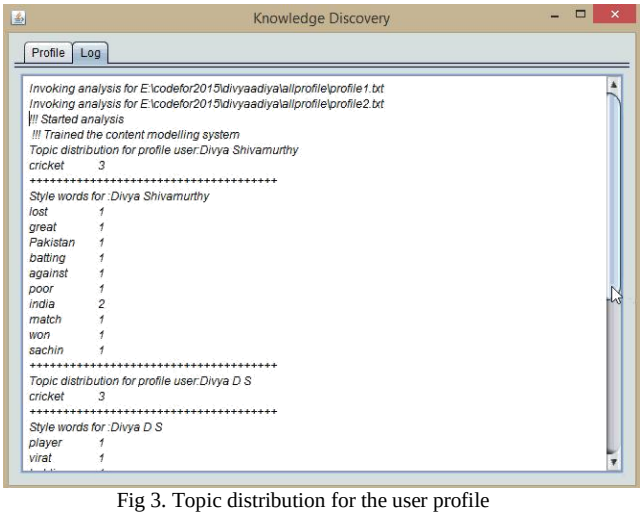

Fig 3. Topic distribution for the user profile

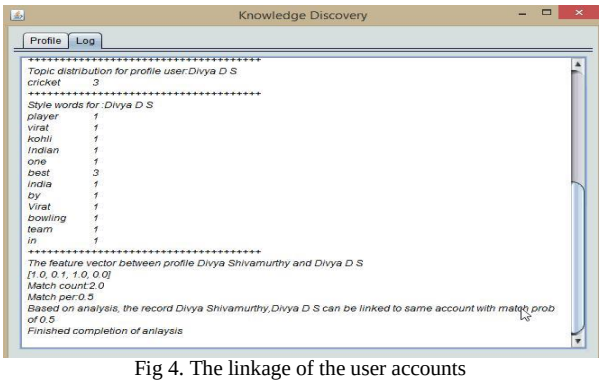

Fig 4. The linkage of the user accounts

The performance of the resulting account linkage is measured along with the parameter time (in milliseconds) and the size of the data. The figure 5 shows the result performance of the two user accounts that is mentioned above.

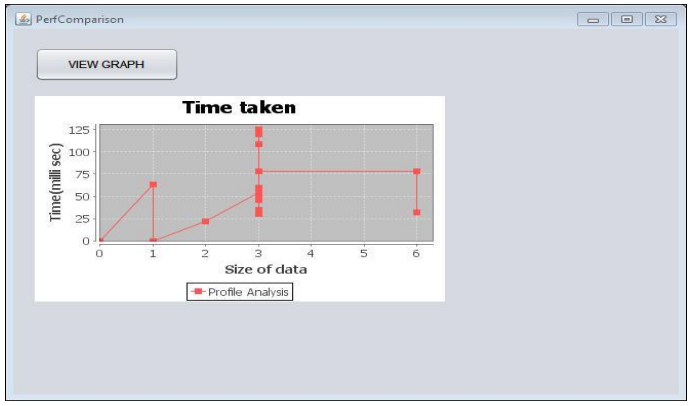

Fig 5. Performance measure

V. CONCLUSION

In this paper, on different social network platforms, we try to link user accounts. To deal with the above mentioned challenges, we propose a framework, called HYDRA, which is a multi-objective learning framework. It incorporates heterogeneous behavior model and also core social network structure. On two real data sets, we tried to evaluate HYDRA against the state-of- the-art. The Experimental results proved that HYDRA performs better than the existing algorithms in identifying true user linkages across different platforms as in figure 5. The data generated in this platform is very large and it cannot be fit into a single PC, it is in the trillions and should be managed by high end servers. In the future, it can be seen that these data are optimized so that minimal servers are being used.

There are a large percentage of missing data in each of the social platforms about 4% in the English social platforms and about 2% in the Chinese social platforms these must be overcome before we can link any accounts. Hence we need a more tight relation with the attributes of all the social platforms, so that this missing information is reduced. With all these available details still the similarities is not more 2% so more aggregation is required for the system to perform better.

REFERENCES

[1] R. Zafarani and H. Liu, "Connecting users across social media sites: A behavioral-modeling approach," in KDD'13, 2013.
[2] 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 WSDM'13, 2013.
[3] T. Iofciu, P. Fankhauser, F. Abel, and K. Bischoff, "Identifying users across social tagging systems," in ICWSM'11, 2011, pp. –1–1.
[4] P. Jain and P. Kumaraguru, "@i to @me: An anatomy of username changing behavior on twitter," CoRR, 2014.
[5] 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.
[6] J. Zhang, X. Kong, and P. S. Yu, "Transferring heterogeneous links across location-based social networks," in WSDM'14, 2014, pp. 303– 312
[7] Y. nan Qian, Y. Hu, J. Cui, Q. Zheng, and Z. Nie, "Combining machine learning and human judgment in author disambiguation," in CIKM'11, 2011, pp. 1241–1246.
[8] D.A Nikam, ,Joshi Govind " Cluster Based Web search", 2012
[9] J. Vosecky, D. Hong, and V. Shen, "User identification across multiple social networks," in NDT'09, 2009, pp. 360–365.