User Profile Matching Analysis across Heterogeneous Online Platforms

MD RAYHANUL MASUD, University of California Riverside, USA

Given two user profiles from heterogeneous online platforms, how can we predict whether they belong to the same user? User profile matching across different platforms has diverse applications including user behavioral analysis, cross-platform monitoring, crime investigation and so on. The task becomes challenging due to the unavailability of user's demographic information, differing interplay inside the user network, privacy issues, and sometimes the intentional impersonation made by the malicious users. Our study presents a comprehensive strategy to measure the similarity between given user profiles analyzing the behavioral footprint available in the user generated contents. The key novelty of our approach is to focus on matching identities across platforms, where users share contents of varied interests and topics manipulating similarity of textual style and temporal activities. We consider Reddit platform as the best suit for our approach for its pool of sub forums called as subreddits to evaluate our results. We show that our approach presents excellent performance in terms of precision, recall and F-1 Score. We claim that user's writing style pattern extracted from their contents and temporal similarity are resourceful features that can be exploited for user profile matching across heterogeneous online platforms for higher accuracy and performance.

CCS Concepts: • **Database Applications** → **Data mining**.

Additional Key Words and Phrases: User Identification, Cross-Media Analysis

ACM Reference Format:
Md Rayhanul Masud. 2021. User Profile Matching Analysis across Heterogeneous Online Platforms. 1, 1 (December 2021), 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In this era of web and technology, people are connected with each other in reality and virtual reality simultaneously. Nowadays, users are open to enormous online platforms of different interests and topics. To get engaged in social networking, users prefer Facebook, Twitter, Instagram whereas, professionals go for LinkedIn, StackOverflow, Reddit. Users produce wealth of information across these platforms. [10] says approximately 42% of online users participate in more than one social platform. As a result, user profiling across these platforms can help us better understand the dynamics of networks. Besides, security professionals can also mine useful investigative information from this profiling to support their works. But nowhere any direct/indirect connection/mapping is maintained. User profile matching thus demands greater interest among the research community in the last decade.

To attempt user matching, lack of reliable and available ground-truth data offers manifold challenges. Because of privacy concerns, platform owners do not share information about their users when requested from law enforcement agencies. On the other hand, users often do not provide their demographic information online publicly to preserve anonymity to the outside world. People produce and consume textual contents in some platforms, and images in some others. These complicate the user matching problem more sophisticated.

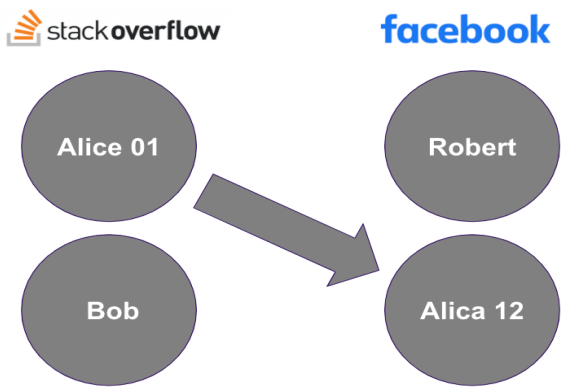

Fig. 1. Alice 01 and Alica 12 are user profiles in StackOverflow and Facebook repectively, though they are maintained by the same user

The problem has been addressed from so many angles: demographic information similarity[4], content similarity[10], image similarity[4], network structural similarity[9] are some to be named. These approaches have mostly focused on the users from online social networking sites though other professional and learning platforms contain different types of content than social ones. In addition, groups of people may not be present across several platforms because of individual choices and interests.

Users can maintain their own way of expression and communication despite their content types may vary. Writing styles and patterns can generate more specific information about a user than the content itself. Furthermore, users usually attend these forums during a specific period of a day/night. Consequently, these behavioral pattern of communication powered by the content similarity can overcome the limitations described above. Featuring this, our approach uses several Machine Learning approaches to match user profiles across four subreddits from Reddit platform. We leverage the common users participating between subreddits as ground truth data. We perform our experimentation on a dataset of 30k posts and comments containing 98 users.

The main contribution of our work is as follows:

- We calculate the similarity score between users based on similarity of writing style
- We calculate the temporal relatedness of user activities

Author's address: Md Rayhanul Masud, mmasu012@gmail.com, University of California Riverside, Riverside, California, USA, 92507.

XXXX-XXXX/2021/12-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

- We model 3 classifiers and evaluate our results based on the ground truth dataset

2 RELATED WORK

User profile matching has been termed as: User matching, User Identification, User Alignment and so on. Related works regarding user profile matching can be divided into 3 following categories:

- `Profile Information Similarity`: When users get registered on a platform, they have to provide some demographic information such as username, gender, date of birth, profile image for identity and so many. Most of the cases, except username, rest of the attributes are not mandatory to be provided. This ensures the privacy of users when participating in platforms if they do not want to share those information willingly. [4] depends on the uniqueness of a username for determining the match. [7], [8] analyzes user behavior in case of selecting usernames and modeled an approach to identify similar users. [5] considers a pairwise comparison model for applying string matching algorithms to find similarity for different profile information.
- `User Generated Content Similarity`: Users share their feelings and expressions with their neighbors in the network through posts, comments and threads. [9] calculates character level and word level cosine similarity for user generated content matching. [2] and [3] propose algorithms to calculate similarity of contents.
- `Network Structure Similarity`: Users in a network interact among them through various types of communication. These interactions include friendship, like-share-comment and follow. [6] depends on the topology of the network to find similarity between users. [1] incorporates the topology along with profile information to model the objective function of user matching.

Profile information attributes are not available always. Some online platforms like FourSquare provides usernames containing only random digits having no significant relatedness with the users. User generated content differs in many degrees from platform to platform. Users and their neighbors may not reside in all platforms collectively and simultaneously. These genuine real life implications add severe limitations to the previous works regarding user matching.

3 PROPOSED METHOD

The proposed method for matching profiles across platforms utilizes the pattern of writing along with textual content, and the temporal pattern of user activities. Our proposed approach exploit three types of features:

- `Pattern of writing`: Users usually either share their own contents or share other contents in the form of url sharing. The proposed method finds the count of url sharing as a metric of similarity. (Figure 2) During matching between users, if the difference of count is lower, the more similar the users are.
- `Temporal pattern of user activities`: When users attend different platforms, they mostly do so at the same period of time during a day. We compare the temporal pattern of posts/comments of the users, and find the similarity score.
- `Textual Content`: Users may share contents of diversified genres. But the pattern of choosing words, punctuation (Figure 3), and paragraphing should be the same unless they themselves try to hide their behavioral footprint. We apply sentence similarity model from https://scikit-learn.org to find the similarity of textual contents.

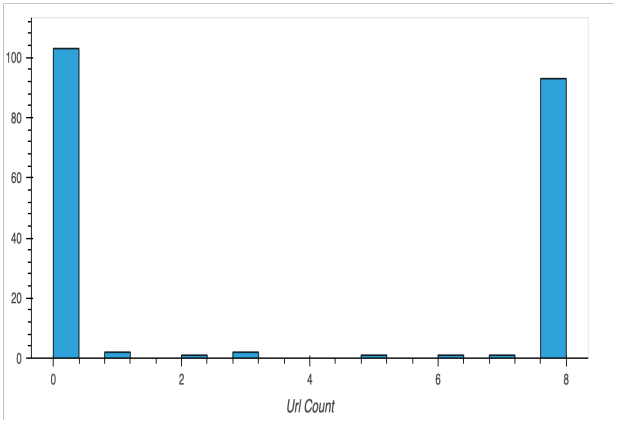

Fig. 2. Histogram of Url Sharing

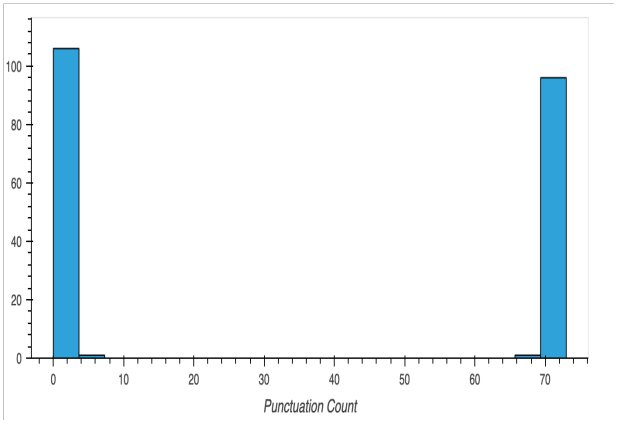

Fig. 3. Histogram of Punctuation Count

We calculate the similarity score for every pair of users in the dataset. The similarity score is defined by the equation below:

$$Score_{sim} = URL_{sim} + TMRL_{sim} + TEX_{sim} \quad (1)$$

where URL_{sim}, $TMRL_{sim}$, TEX_{sim} and $Score_{sim}$ are the score of matching writing pattern, the score of matching temporal pattern of activities, the score of textual similarity, and the total similarity respectively. When measuring the score, we consider all of the posts and comments of a single user for a specific platform to make a distinct profile. All the scores are then populated in a similarity score matrix where each row represents a single user profile in a

platform. Each column provides the similarity of that user with the other ones in the dataset.

4 EXPERIMENTAL EVALUATION

4.1 Dataset:

We use Reddit Platform for applying our methods to do experimentation. Since users do not need to register different user accounts for participating in different subreddits, a good number of users overlap across subreddits. To make the best use of this advantage, we have identified 4 subreddits (Table 1) where the maximum overlap of users across subreddits is quiet high. We also ensure the fact that the subreddits need to be dissimilar with each other. These subreddits will resemble heterogeneous online platforms for our experimentation.

SubReddit	Users	Posts
datascience	23	6112
learnprogrammer	32	5419
learnpython	24	3205
programmerhumor	19	17185

Table 1. Statistics of dataset

4.2 Ground Truth Data Generation:

We identify several top users of those subreddits. With the help of an online tool https://github.com/JakapunTachaiya/reddit_praw, we crawl 1094 threads including 62655 comments of 98 users across the platforms. As the users have the same identity for every subreddit, we have labeled them accordingly.

4.3 Evaluation Metrics

a. Precision: Precision is the percentage of true matching pairs among the predicted matching pairs.

$$precision = \frac{TP}{TP + FP} \tag{2}$$

b. Recall: Recall is the percentage of predicted true matching pairs among the true matching pairs.

$$precision = \frac{TP}{TP + FN} \tag{3}$$

c. F-1 Score: The F1 score is defined as the harmonic mean of precision and recall.

$$precision = \frac{2 * precision * recall}{precision + recall} \tag{4}$$

4.4 Evaluation

We train three ML classifiers (Bagging, Random Forest and Adaboost) separately to do our experimentation. We consider Precision, Recall and F1-Score for metrics of performance evaluation of our approach. Accuracy alone might not judge the performance. So precision and recall offer greater significance in any analysis. Table 2 depicts that Bagging works relatively better than other 2 classifiers

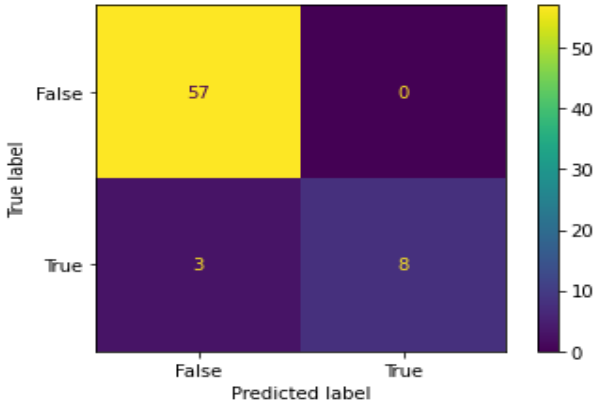

Fig. 4. Confusion Matrix for Bagging Classifier

in terms of Precision, Recall and F1-Score. All of the classifiers perform quite well. The precision is over 80% for each of them which justify the algorithm that we propose.

Method	Class	Precision	Recall	F1-Score
Bagging Classifier	True	100%	73%	84%
	False	95%	100%	97%
Random Forest Classifier	True	100%	64%	78%
	False	93%	100%	97%
Adaboost Classifer	True	83%	45%	59%
	False	90%	98%	94%

Table 2. Performance of different models

Figure 4 shows the confusion matrix for Bagging Classifier. All the non-matching pairs have been properly detected by the classifier, though the model miss a few of matching pairs. This may happen as one of the profiles from a pair might have relatively less information than the other to measure the similarity among them.

5 DISCUSSION AND CONCLUSIONS

User profile matching problem has manifold challenges to be addressed. Lack of ground-truth dataset, varied nature of contents across platforms, peculiarity of human nature - all have turned the problem to be more challenging. In this work, we show how writing styles and temporal similarity of activities can help user matching when contents across platforms are very dissimilar. We find users of reddit platform often share links of other websites/blogs. The frequency of sharing url for the same users in subreddits is found very similar. Furthermore our result shows that users participate in online platforms regularly during a specific period of the day. These behavioral patterns are found to remain the same across heterogeneous platforms. We leverage these features, and feed into the machine learning models to accomplish our goals. The results of the classifiers show satisfactory performance. We believe that our models will work across heterogeneous platforms with high performance and accuracy.

REFERENCES

[1] Sergey Bartunov, Anton Korshunov, Seung-Taek Park, Wonho Ryu, and Hyungdong Lee. 2012. Joint link-attribute user identity resolution in online social networks. In *Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining, Workshop on Social Network Mining and Analysis. ACM.*

[2] Xiangnan Kong, Jiawei Zhang, and Philip S Yu. 2013. Inferring anchor links across multiple heterogeneous social networks. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management.* 179–188.

[3] Yongjun Li, Zhen Zhang, You Peng, Hongzhi Yin, and Quanqing Xu. 2018. Matching user accounts based on user generated content across social networks. *Future Generation Computer Systems* 83 (2018), 104–115.

[4] Jing Liu, Fan Zhang, Xinying Song, Young-In Song, Chin-Yew Lin, and Hsiao-Wuen Hon. 2013. 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.* 495–504.

[5] Ravita Mishra. 2019. Entity resolution in online multiple social networks (@ Facebook and LinkedIn). In *Emerging Technologies in Data Mining and Information Security.* Springer, 221–237.

[6] Arvind Narayanan and Vitaly Shmatikov. 2009. De-anonymizing social networks. In *2009 30th IEEE symposium on security and privacy.* IEEE, 173–187.

[7] Reza Zafarani and Huan Liu. 2009. Connecting corresponding identities across communities. In *Third International AAAI Conference on Weblogs and Social Media.*

[8] Reza Zafarani and Huan Liu. 2013. Connecting users across social media sites: a behavioral-modeling approach. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining.* 41–49.

[9] Jing Zhang, Bo Chen, Xianming Wang, Hong Chen, Cuiping Li, Fengmei Jin, Guojie Song, and Yutao Zhang. 2018. Mego2vec: Embedding matched ego networks for user alignment across social networks. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management.* 327–336.

[10] Yutao Zhang, Jie Tang, Zhilin Yang, Jian Pei, and Philip S Yu. 2015. Cosnet: Connecting heterogeneous social networks with local and global consistency. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining.* 1485–1494.