COMP530 - Project Proposal
Profile Matching

Erhan Tezcan · Waris Gill · Mandana Bagheri Marzijarani
etezcan19@ku.edu.tr · wgill18@ku.edu.tr · mmarzijarani20@ku.edu.tr

03.11.2020

1 Introduction

Over the past decade many Online Social Networks (OSNs) have emerged and each offer a diverse set of features/services. Many users have accounts on multiple OSNs. It is reported that the number of social network accounts per person is 8.6 [7]. Popularity of users with accounts on multiple OSNs has created a lot of opportunity for profile matching research which is aimed at finding users with the same identity across multiple OSNs using state of the art Machine Learning models [3]. Profile matching provides complete insight over individuals to online service builders in a way that is not possible by analyzing them on a single OSN's platform.

From a security and privacy standpoint profile matching can represent a serious privacy concern. If an attacker can match anonymous profiles of individuals who consciously hide their real identity on one platform, to their real identities in another OSN, she can put together sensitive information about individuals that are meant to be private. The importance of profile matching research lies in the fact that it can quantify and mitigate the risk of such privacy attacks and help OSNs harden their policies to safeguard their user's privacy. In this research we aim at building a Machine Learning (ML) model for profile matching that improves over previous efforts.

2 Problem Statement

Given the variety of social networks, the users create accounts in different ways as the rules and regulations are different from one another. However, almost all of these have a few things in common: i.e. every user has a username, some content, and friends as well as followers. The attributes like username, content, and followers will vary from one social media platform to another. So it is quite challenging to match a single user's profiles among many OSN's, as there is no unique global identifier (e.g.: username) for the user. This problem is defined as: "Identity Resolution in Online Social Networks (IROSNs)". IROSNs can be defined as: "Given the user attributes (e.g.: name, username, location, content etc.) in a social networking site, identify him/her on another social network platform". For instance, given a user's Facebook profile, find his/her Twitter profile. IROSNs techniques has the following applications:

1. **Protecting User Privacy:** IROSNs tools can protect/increase the privacy of a user. On some social media sites, the users hide their personal

information and just share their personal views (about religion, politics, etc.) while on some they have some public information such as username, region, post time, followers who may have similar views, but they are posting them publicly. So, in such cases, an attacker can easily learn who is the anonymized user which is very critical and dangerous in some scenarios. Thus, it is very important to develop a privacy tool which can quantized the anonymity level of user profile and suggest ways to increase its privacy.

2. **Identifying Malicious Users:** With the help of IROSNs, malicious users can be identified on multiple platforms. The malicious users obfuscate their attributes so that they cannot be detected on other platforms. So, behavioural-based IROSNs can be utilized to find and discard such users on online social networks.

3. **Recommendation Systems:** IROSNs can also suggest friends based on the information of one network on another. Additionally, it can also suggest content such as music, videos, and articles based on the user's interests and posts.

3 Technical Details

Our focus is the two social media platforms: Twitter and Facebook. We will try to match a Twitter user to a Facebook user, and likewise a Facebook user to a Twitter user. With this in mind, we will have several approaches to consider, both on their own and all-together:

- We will use their public information in doing so, such as username, first name, last name, birth date, gender, profile image (if exists).
- We will check several latest tweets for information, which might be useful in cases such as when the user shares a post on many accounts simultaneously or shares their Facebook post in Twitter.
- We will process the user network (i.e. followers, followees). It is certainly possible that a Twitter user may not be matched to a Facebook user, however if we could easily match Facebook accounts of several followers or followees, then we can perhaps find candidates by looking at their friends on Facebook.

We will try to improve upon existing works, by implementing more advanced algorithms and trying to find different matching factors. For example, the 2013 works of [2] uses a standard RGB histogram matching for profile image similarities, whereas the 2020 works of [1] uses OpenFace [6] to do face recognition. Similarly, the prior works used straightforward syntactic matching of text content, whereas modern works employ advanced natural language processing methods such as sentiment analysis for content matching. We believe there is a room for improvement in similarity measurements, especially with the more narrowed down version of IROSN where we focus on Twitter and Facebook alone. With these in mind, we have decided to use Python language for our implementations, given it's ease of use, performance, and great variety of Machine Learning libraries.

3.1 Dataset

Since the project requires a dataset of social network users, we believe it should explained in detail. We plan to build our dataset from an equal number of Facebook and Twitter users and their publicly available information that we can use as features in our ML model. Our goal is for this dataset is to include both matchable and unmatchable pairs of Twitter and Facebook profiles. The challenge in creating such dataset is that given a set of Facebook and Twitter profiles, finding a subset consisting of all matchable profiles (i.e. users who have the same identity) as ground truth is difficult. One way to find the ground truth subset is to manually compare Facebook and Twitter users against each other and determine the ground truth subset, however, this is not a feasible option given the huge number of users required to be included in such research. The other option is to build a dataset of only matchable profiles and base the correctness of our model on the ratio of the matched profiles our model can detect, but this method is prone to overfitting and would fail to report false positives that are an important part of evaluation of such a model in real world [5].

To build a dataset that includes both matchable and unmatchable profiles we can intuitively think of choosing random Twitter and Facebook profiles but the problem with this idea is that first the aforementioned problem with finding the ground truth still stands, and second the chance that no matchable pair in such database exists is high given the fact that both platforms have hundreds of millions of users but we would at most examine a very small subset of them. To mimic a real world dataset, we decided to collect our data in a way that will consist of two subsets, up to 10 percent of matchable profiles (our ground truth subset) and 90+ percent of unmatchable profiles. This decision is motivated by the works of [2, 3, 5].

To do so we start by looking for profiles on both platforms that includes the link to the other platform or have the same username, then we manually confirm that they are matched profiles and include them in our ground truth dataset. We then choose the rest of our dataset aiming at finding profiles across platforms that won't be matched, though of course we cannot be 100 percent sure that we can come up with completely unmatchable profiles, but if we choose the profiles in some random way we can minimize the risk, for example we can choose two public figures from two different fields like politics and network security, and include their followers in our dataset.

4 Evaluation Criteria

As described in section 3, we will be creating the dataset ourselves. We have already built a database hosted on a Vultr [8]server so that all group members can work on it. We will query both Facebook graph [9] and Twitter developer's API [10] to find publicly available information and even recent content published by the users. We will parse the query's response and store it in SQL tables in our database, we would also keep raw response in JSON format to be able to extract more features from it later just in case we need them. Using the extracted information we will build a database of information for our Facebook and Twitter profiles. Naturally, there are limitations on how to collect our

required data from both APIs. Twitter's developers API sets a limit of 30 queries per day and Facebook graph API has a 200 requests per hour limit. Given the limit on the number of queries and also the number of users that have been included in recent research datasets, we think it makes sense to create a dataset of 1000 Facebook users and 1000 Twitter users including 10 percent matchable profiles.

Our main performance metrics are Precision, Recall and Accuracy, as defined in Appendix A. We chose these metrics, Precision and Recall especially, because we have seen that accuracy alone might not be enough, as mentioned in [1]. With respect to the size of our dataset, if possible, we may have an overall performance evaluation by using Identity-based Accuracy, described in [4], and shown in Appendix A.

5 Proposed Deliverables

- In terms of functionality, we will deliver a Python script, that takes a Twitter or Facebook username as input, and matches to a set of candidate users in the dataset from it's respective social network to the other. The results will be shown in an HTML file, for clarity and better UX.
- A final report will be delivered, which includes the performances of implemented algorithms, with metrics such as Accuracy, Precision and Recall and examples of matched users and how they were matched.
- Our source code with clear instructions and documentation will be delivered.

6 Tentative Timeline

	W5	W6	W7	W8	W9	W10	W11	W12	W13	W14
Preparing Dataset	M									
Preparing Pipeline		M, W, E								
Implementing Algorithms			W, E		W, M, E					
Evaluations						W, M, E				
UX Module									E	

Figure 1: GANTT Chart.

In figure 1 we present our GANTT chart with respect to the 10 weeks of this project. Dataset is the top priority, and though we will of course be working as a group, Mandana (shown as M) will be primarily responsible with setting up the database. Then, the overall structure of the algorithm will be created as an interface, primarily by Erhan (E) and Waris (W), while Mandana is continuing with the dataset. We will then implement the algorithms, which is the heart of this project. We will first look at the user and content based methods, and then move into graph methods. Our evaluations and UX follows next, however keep in mind that algorithms are subject to change and refactor during the course of project. UX module not only includes the output HTML itself, but also

enables the user to allow several parameters during the search (e.g. Similarity Threshold).

References

[1] Halimi, A., & Ayday, E. (2017). Profile Matching Across Unstructured Online Social Networks: Threats and Countermeasures. ArXiv, abs/1711.01815.

[2] Jain, P., Kumaraguru, P., & Joshi, A. (2013). @i seek 'fb.me': identifying users across multiple online social networks. WWW '13 Companion.

[3] Shu, Kai & Wang, Suhang & Tang, Jiliang & Zafarani, Reza & Liu, Huan. (2017). User Identity Linkage across Online Social Networks: A Review. SIGKDD Explorations. 18. 10.1145/3068777.3068781.

[4] Zhang, H., Kan, M., Liu, Y., & Ma, S. (2014). Online Social Network Profile Linkage. AIRS.

[5] Oana Goga, Patrick Loiseau, Robin Sommer, Renata Teixeira, and Krishna P Gummadi. (2015). On the reliability of profile matching across large online social networks. In KDD

[6] Amos, B., Ludwiczuk, B., Satyanarayanan, M. (2016). Openface: A general-purpose face recognition library with mobile applications. Tech. rep., CMU-CS-16-118, CMU School of Computer Science

[7] https://backlinko.com/social-media-users *Accessed 03.11.2020*

[8] https://www.vultr.com/ *Accessed 03.11.2020*

[9] https://developers.facebook.com/docs/graph-api/ *Accessed 03.11.2020*

[10] https://developer.twitter.com/en *Accessed 03.11.2020*

Appendices

A. Metrics

- Accuracy[1]:
$$\frac{|TP| + |TN|}{|TP| + |TN| + |FP| + |FN|}$$
- Precision:
$$\frac{|TP|}{|TP| + |FP|}$$
- Recall:
$$\frac{|TP|}{|TP| + |FP|}$$

[1] TP stands for True Positive, TN stands for True Negative, FP stands for False Positive and FN stands for False Negative.

- Identity Based Accuracy (I_{acc})

$$I_{acc} = \frac{\text{\# correctly identified user identities}}{\text{\# ground truth user identities}}$$