Brave New Task: User Account Matching

Claudia Hauff*
Delft University of Technology
Delft, The Netherlands
c.hauff@tudelft.nl

Gerald Friedland†
International Computer Science Institute
Berkeley, USA
fractor@icsi.berkeley.edu

ABSTRACT

Today, the usage of online social networks and their many manifestations is widespread. Users tend not to be just active on a single platform, they take advantage of a range of platforms for different purposes: Twitter for microblogging, Flickr for sharing pictures, YouTube for sharing videos, etc. The question we consider in this task is to what extent users (unintentionally) leak information in their social Web streams, either through their direct contributions or the meta-data of their contributions, that allows us to automatically identify and match their corresponding accounts *across* streams. In this paper, we present the data set we developed for the user account matching task and the baseline results.

1. INTRODUCTION

Users who are active on the social Web produce digital traces continuously by posting messages, sharing videos, commenting on news items, or simply by walking into a store (and checking in on Foursquare). The number of traces a user generates per day is steadily increasing thanks to the myriad of online social networks and the widespread use of smart phone apps which often publish messages on behalf of the user in a semi-automatic manner.

In this task, we consider the question to what extend we can exploit these traces to automatically link accounts on different social Web streams to the same user. While some users may intentionally want to be recognized across social networks (by using the same user name and/or posting links to their various social Web profiles), others may want to keep their connections private and may not even be aware about the amount of information they leak that make them identifiable.

Previous work in this direction has mostly focused on information available directly in the user profiles. Zarani and Liu [7] conducted a large-scale study on identifying users across communities which is based exclusively on the similarity between user names on different platforms. The reported accuracy level indicates that such a simple approach can be successful for *cooperative* users, i.e. users that use similar identifiers across platforms.

To what extent a user profile, that has been aggregated across social networks, is richer in information than the respective user profiles of each individual social Web platform has also been investigated in a number of studies, e.g. [6, 4, 1, 2]. In all instances, the ground truth is derived by crawling the public profiles available in online identity management portals such as the (now defunct) Google Social Graph API. In [4] a password recovery attack based on such information has been described while in [2] it has been shown how such enriched profiles can be exploited to gather additional personal data about users that is not available online. Exploitation of the graph structure of social networks (instead of user profile information) for de-anonymization attacks has also been shown to yield good results in the past [5]. Finally, Iofciu et al. [3] went beyond user matching based on profile information and included content-based information (tags users assigned to images and bookmarks) when matching Flickr, Delicious and StumbleUpon user accounts. Compared to user name based matching, they found content based matching to be much more difficult. This result is the starting point for our task: assuming a set of *uncooperative* users, i.e. users that cannot be linked according to their self-reported profile information, to what extent is it still possible to determine likely matches? In year one of this task, we consider only two social Web streams: Flickr and Twitter. We formally define the task as follows: *Given the Flickr stream F_{u_i} of user u_i, and a set of N Twitter streams $\mathbf{T} = \{T_1, T_2, ..., T_N\}$, rank the streams according to their likelihood of having been produced by user u_i.*

Ideally, the stream T_{u_i}, that has been produced by u_i appears at the top of the result ranking.

2. DATA SET

For a period of three months we followed approximately 50,000 randomly chosen users that themselves followed various mainstream political parties in the Netherlands, Great Britain and Germany. This setup yielded a very diverse set of users in terms of languages used, the type of users (individuals vs. organizations vs. businesses) and the amount of activity in the stream. A total of $N = 18,372$ of these users posted at least one message during this time period and thus their Twitter streams form the set $\mathbf{T}$. We then determined

*This research has received funding from the European Union Seventh Framework Programme (FP7/2007-2013), grant agreement no ICT 257831 (ImREAL project).
†This material is based upon work supported by the National Science Foundation under Grant No. CNS-1065240. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

MediaEval 2012 Workshop, October 4-5, 2012, Pisa, Italy

a set of matching Flickr accounts in a two-step procedure[1]:

1. We identified potential matches (i) by searching in **T** for tweets with URLs containing `flickr.com` and (ii) by querying the Flickr API with each of our N Twitter user names for accounts with the same name on Flickr.
2. Each of the potential matches was manually verified by considering the information available in the user profiles as well as the stream content. Potential matches that could not be verified with a high degree of confidence were ignored.

In total we found 233 verified Twitter-Flickr account matches. They form our ground truth. For each match, we crawled the respective Flickr account; Fig. 1 shows a scatter plot of the number of Flickr images vs. the number of Twitter messages for each account pair. The average number of images is 1053 (median 200), while the average number of Twitter messages per account is 921 (median 290). Fig. 2 visualizes the temporal distribution of the Flickr images (the date they were taken) and Twitter message (the date they were posted). While the Flickr API allows unrestricted access to posted items with respect to time, Twitter has a very restrictive policy and thus the tweets in our data set are restricted to a three month period.

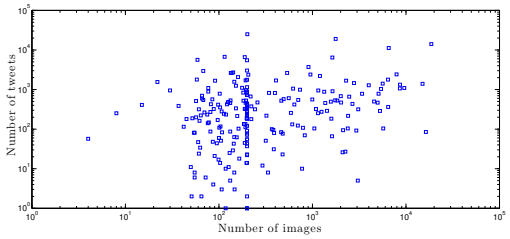

Figure 1: Scatter plot of the number of Flickr images and Twitter messages for the 233 matching accounts.

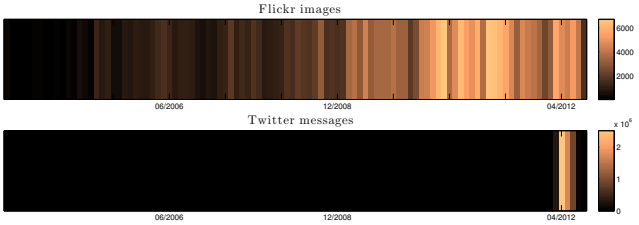

Figure 2: Temporal distribution (in months) of the number of Flickr images and Twitter messages for the 233 matching accounts. Time period: 01/2004 to 08/2012.

The final data set statistics are shown in Tab. 1; here, we distinguish the matching account pairs (ground truth) and the set **T**. Recall that we use the information in each Flickr stream to find the correct corresponding Twitter stream within the approximately 18,000 Twitter streams.

3. BASELINE RESULTS

To provide a baseline, we treat the task as an information retrieval problem: the textual information of all Flickr images by a user are concatenated and treated as a *query*. All tweets by a single user are concatenated and treated as a *document*. We then rank the $\approx 18,000$ documents (i.e. the set of streams **T**) for each query (Flickr account) according to a standard retrieval approach (Okapi). In line with previous work [3] we evaluate the quality of the matching algorithm in mean reciprocal rank (MRR). The results in Tab. 2 confirm the difficulty of the task: the MRR is low and for more than 60% of our matched accounts the reciprocal rank is zero. A correlation analysis shows that the number of items (Flickr images and Twitter messages) can only explain the results to a small degree - the number of Flickr images available is moderately positively correlated with MRR.

[1]We considered the use of an online identity aggregation service to determine a set of matching accounts as done in previous work. In preliminary experiments however, we found that a significant number of users list accounts of friends or famous personalities in their profiles.

Number of account pairs	233
Number of tweets in account pairs	214,664
Number of images in account pairs	245,320
Number of Twitter accounts N	18,372
Number of tweets in **T**	2,795,388

Table 1: Data set statistics of the task.

MRR	$\tau_{\#tweets}$	$\tau_{\#images}$	$\tau_{\#tweets+\#images}$
0.168	0.033	0.200	0.146

Table 2: Results of the Okapi baseline. Columns 2-4 show the rank correlation (Kendall's τ) between MRR and the number of images and tweets available.

4. CHALLENGES

Lastly, we present a number of challenges we encountered during the setup of this task:

- Social networks often place a limit on the amount of data that is publicly accessible and a long-term experimental setup is required to gather a large amount of data.
- A user may use different social networks at different time periods - matching a user who is currently active on Twitter, to his Flickr account that was last used two or three years ago is difficult.
- A considerable number of the encountered matched accounts were not operated by private individuals, but belong to organizations or business endeavours.
- Automatic or semi-automatic methods to generate pairs of matched accounts are not always reliable. In particular, matching users through self-reported links in online identity management services has a non-negligible error rate.
- Implicitly, the users we selected were cooperative as we were able to manually match them according to their profile information, avatar image or content. How to obtain a set of uncooperative users is an open question.

5. REFERENCES

[1] F. Abel, N. Henze, E. Herder, and D. Krause. Interweaving public user profiles on the web. In *UMAP '10*, pages 16–27, 2010.

[2] T. Chen, M. A. Kaafar, A. Friedman, and R. Boreli. Is more always merrier?: a deep dive into online social footprints. In *WOSN '12*, pages 67–72, 2012.

[3] T. Iofciu, P. Fankhauser, F. Abel, and K. Bischoff. Identifying users across social tagging systems. In *ICWSM '11*, 2011.

[4] D. Irani, S. Webb, K. Li, and C. Pu. Modeling unintended personal-information leakage from multiple online social networks. *IEEE Internet Computing*, 15(3):13–19, 2011.

[5] A. Narayanan and V. Shmatikov. De-anonymizing social networks. In *IEEE Symposium on Security and Privacy*, pages 173–187, 2009.

[6] J. Vosecky and D. H. V. Y. Shen. User identification across multiple social networks. In *NDT '09*, pages 360–365, 2009.

[7] R. Zafarani and H. Liu. Connecting corresponding identities across communities. In *ICWSM '09*, 2009.