personaLink: A tool to Link Users across Online Forums.

Chirag Nagpal, Benedikt Boecking, Kyle Miller, Artur Dubrawski

August 25, 2016

Abstract

Most online internet forums, follow a relatively simple, and similar schema, with only a little change in the number of features associated with a particular user profile. In this paper, we describe, personaLink, a tool that exploits machine learning to provide a pairwise probability of user profiles representing the same entity across two different online forums. personaLink, is a complete end-to-end, python based program, that treats the aforementioned identity resolution task as a binary classification problem.

1 Introduction

Entity Resolution is an age old database management problem. In todays day and age, with the advent of multiple social media platforms, the challenge is more relevant. Entity Resolution has been extensively studied in the past.[1, 2, 3, 4, 5] Various studies have provided theoretical performance guarantees of Entity Resolution results[6, 7]. Entity Resolution has also been used to tackle social and law enforcement issues like Human Trafficking[8, 9].

While, there are multiple social networking sites, forums, newsgroups etc., most of them follow similar schemas, with little or no difference. For example, most websites, have the field `'username'`, which is a unique handle associated with a user, `'registrationTime'` which represents the timestamp the profile was created etc.

personaLink, aims to leverage this common structure, in order to perform entity resolution ,between two different platforms. The tool requires a small amount of ground truth as training data to begin, It then performs feature extraction to generate data for the classification task, and fits an ensemble model. These steps are transparently abstracted away from the end user, making it easy for the user to employ the tool, with out having to individually perform each task.

The rest of this paper is organised as follows, the Featurisation section deals with the kind of features extracted, the Classification section deals with the ensemble methods utilised and there sanity for the task at hand. The Usage

section deals with how personaLink can be employed from an enduser perspetive. We also provide results of personaLink on the MEMEX summer CP2 task.

2 Featurisation

This section deals with the specific features extracted by personaLink.

2.1 Username

Username is a very strong feature, in the MEMEX classification task, exact username match had a pairwise precision of around 80%. Usernames have also been studied extensively in the past [10]. We extract multiple features from pairs of usernames as given in Table 2.

1	Levenshtein Distance Between Usernames
2	Levenshtein Distance Between Usernames (Normalised)
3	Levenshtein Distance Between Lowercase Usernames
4	Levenshtein Distance Between Lowercase Usernames (Normalised)

Table 1: Features extracted from Username

Site A	Site B
`Ivan-1`	`Ivan999`
`NoxJoker0`	`noxjoker`
`DimaSS_55`	`DimaSss`
`Delivery M`	`Delivery Man`

Table 2: Exaples of username similarity.

2.2 Registration Time

The absolute difference between the users' registration time on a forum, most datasets give this value in the form of a UNIX timestamp.

$$|T_a - T_b|$$

2.3 Time Zone

Time Zone can be an informative signal can give a rough estimate of the users location. Most forums have the Time Zone of the user encoded in UTC format, exact match between Time Zones is used as a feature.

2.4 Group Affliations

Forums allow users to be part of certain groups of users with common interest, some groups memberships are voluntary, while some are regulated by 'community leaders' or 'moderators'.

personaLink encodes this information in the feature vector as the number of groups both the users are a part of. Certain preprocessing like conversion to lowercase is performed to increase probability of match between similar groups.

Table 3 illustrates some positive pairs with similar group affliations.

Site A	Groups	Site B	Groups
Antilimit	Rippers	antilimit	RIPPER
Shout777	Banned	SHOUT777	Banned
max1006	Registered, Old	max1006	Registered

Table 3: Positive Pairs with similar Group affliations

2.5 Instant Messaging IDs

Forums display users IM IDs, like Jabber, Skype, AOL ID etc. personaLink collapses all the user IDs corresponding to the user in a single set, and uses then the cardinality of the union of these sets as a feature.

2.6 Textual Features

While, most profile pairs would be dependent on the username to perform resolution, the strength of personaLink lies with its ability to also incorporate textual information from a users Posts on the forum to predict a match.

Simple text similarity metrics like Cosine and Jaccards similarity performs well and are extracted. personaLink calculates individual pairwise Text similarity between each pair of a users post, and aggregates like the minimum, average and standard deviation of the pairwise distance.

Text similarity between users, by collapsing every user to a single document are also extracted. Table 4 lists the fetures extracted by personaLink.

1	Max Cosine Similarity of each pair of posts
2	Min Cosine Similarity of each pair of posts
3	Mean Cosine Similarity of each pair of posts
4	Max Jaccards Similarity of each pair of posts
5	Min Jaccards Similarity of each pair of posts
6	Mean Jaccards Similarity of each pair of posts
7	Cosine Similarity between each user
8	Jaccards Similarity between each user

Table 4: Textual Features from Posts

2.7 Temporal Features

Most forums have a timestamp associated with a users posts, this temporal data can contain some information about a users posting patterns and trends. The temporal features extracted are given in 5

1	Absolute difference between Posts Mean, $\lvert\mu_A - \mu_B\rvert$
2	Absolute difference between Posts Std Dev, $\lvert\sigma_A - \sigma_B\rvert$

Table 5: Temporal Features from Posts

3 Classification

We treat the task as a binary classification problem, and experiment with a few different Ensemble methods. Ensemble methods have been shown to outperform other models in most classification tasks [11, 12]. We use Scikit-Learn [13] and experiment with Random Forests, AdaBoost [14] (with a Decision Tree Base Estimator) and Gradient Boosted Regression Trees [15, 16] and evaluate the performance in 3 fold cross validation over the given dataset.

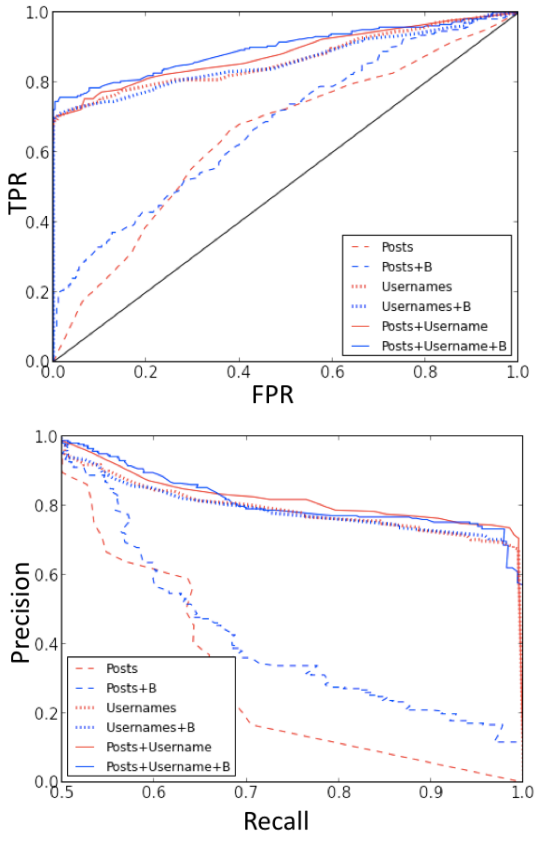

Figure 1: ROC curves for the training of the classifierss.

GBRT does outperforms Adaboost and ordinary Random Forests on the given data ($p \sim 0.049$) and hence is a natural choice for the classification task. It is also apparent from the 'bump' of the ROC in the high precision (FPR < 0.1) that a reasoable amount of pairs are correctly labelled, just with the features extracted from the posts.

4 Results

We present results for personaLink on the MEMEX Summer Challenge Problem 2 Dataset. The ground truth provided consisted of 565 positive pairs of users known to be matched. We trained personaLink over all instances of negative pairs extracted exhaustively from the given ground truth. Inorder to ensure class balance, we duplicate the data from the postive pairs.

The evaluation dataset consisted of 1347 positive samples. The results for the tasks are provided as ROC plots given below.

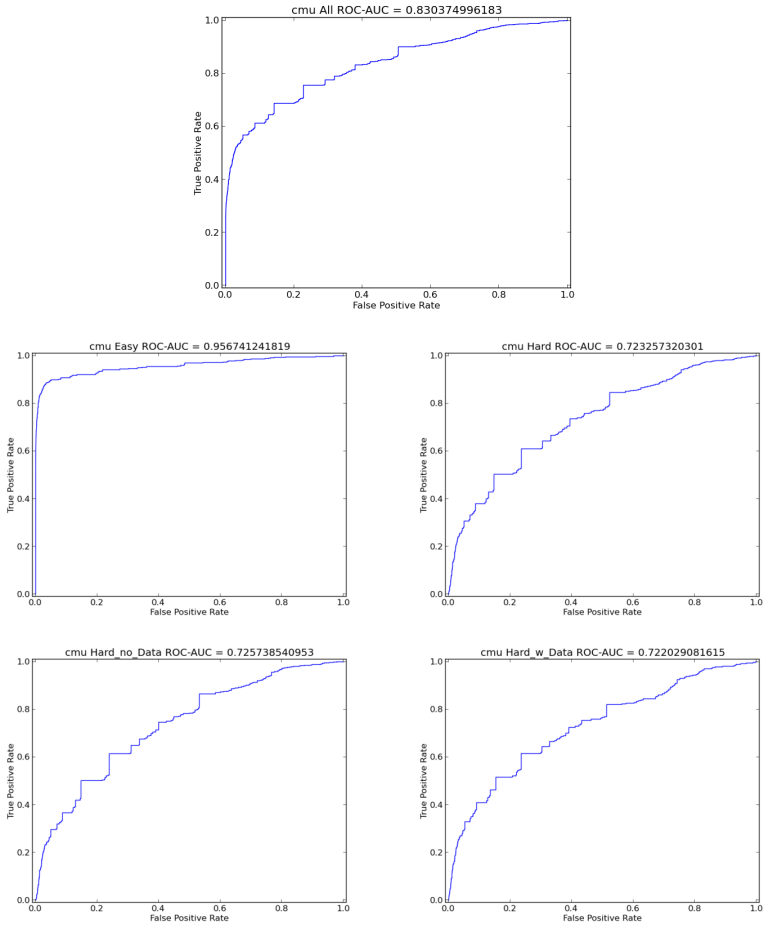

Figure 2: ROC curves for the evalution of the MEMEX Summer CP2 Challenge

5 Usage

All the code for personaLink is opensource and available on github at .

personaLink outputs the featurised data in the form of a comma separated file which can then be utilised to fit a classifier.

5.1 Generating Training Data

```
$ python personaLink.py train <folder with data> <output file name>
```

5.2 Generating Test Data

```
$ python personaLink.py test <folder with data> <output file name>
```

References

[1] Tereza Iofciu, Peter Fankhauser, Fabian Abel, and Kerstin Bischoff. Identifying users across social tagging systems. In *ICWSM*, 2011.

[2] Mine Spears. Method and system for matching users for relationships using a discussion based approach, June 13 2005. US Patent App. 11/151,473.

[3] Oana Goga, Patrick Loiseau, Robin Sommer, Renata Teixeira, and Krishna P Gummadi. On the reliability of profile matching across large online social networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1799–1808. ACM, 2015.

[4] Olga Peled, Michael Fire, Lior Rokach, and Yuval Elovici. Entity matching in online social networks. In *Social Computing (SocialCom), 2013 International Conference on*, pages 339–344. IEEE, 2013.

[5] Haochen Zhang, Min-Yen Kan, Yiqun Liu, and Shaoping Ma. Online social network profile linkage. In *Asia Information Retrieval Symposium*, pages 197–208. Springer, 2014.

[6] Matt Barnes, Kyle Miller, and Artur Dubrawski. Performance bounds for pairwise entity resolution. *arXiv preprint arXiv:1509.03302*, 2015.

[7] Omar Benjelloun, Hector Garcia-Molina, David Menestrina, Qi Su, Steven Euijong Whang, and Jennifer Widom. Swoosh: a generic approach to entity resolution. *The VLDB JournalThe International Journal on Very Large Data Bases*, 18(1):255–276, 2009.

[8] Chirag Nagpal, Kyle Miller, Benedikt Boecking, and Artur Dubrawski. An entity resolution approach to isolate instances of human trafficking online. *arXiv preprint arXiv:1509.06659*, 2015.

[9] Artur Dubrawski, Kyle Miller, Matthew Barnes, Benedikt Boecking, and Emily Kennedy. Leveraging publicly available data to discern patterns of human-trafficking activity. *Journal of Human Trafficking*, 1(1):65–85, 2015.

[10] Daniele Perito, Claude Castelluccia, Mohamed Ali Kaafar, and Pere Manils. How unique and traceable are usernames? In *International Symposium on Privacy Enhancing Technologies Symposium*, pages 1–17. Springer, 2011.

[11] Antti Puurula, Jesse Read, and Albert Bifet. Kaggle lshtc4 winning solution. *arXiv preprint arXiv:1405.0546*, 2014.

[12] Mario Lasseck. Bird song classification in field recordings: winning solution for nips4b 2013 competition. In *Proc. of int. symp. Neural Information Scaled for Bioacoustics, sabiod. org/nips4b, joint to NIPS, Nevada*, pages 176–181, 2013.

[13] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011.

[14] Yoav Freund and Robert E Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*, pages 23–37. Springer, 1995.

[15] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[16] Jerome H Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378, 2002.