De-anonymizing Users Across Heterogeneous Social Computing Platforms

Mohammed Korayem and David J. Crandall
School of Informatics & Computing, Indiana University, Bloomington, IN, USA

1. Motivation

- People using multiple social networking websites may be willing to share personal details on some sites but not on others. For instance, they may reveal their name on Facebook but prefer to stay anonymous on Twitter.
- But can correlations in activity across the sites be used to infer that two accounts correspond to the same person?
- We study how similarity of temporal, textual, geospatial, and social features of users' publicly-accessible activity data can be used to link accounts across sites.

2. Data

- We collected Flickr photos and Twitter tweets between May 1, 2010 and August 31, 2010 using public APIs.
- We build ground truth by finding pairs of accounts (probably) corresponding to the same user by looking for identical usernames and hometowns.
- Dataset includes 108,206 tweets and 589,045 photos by these 3,538 matching people during the four-month period.

3. Classifiers

- We take a machine learning approach, learning classifiers that decide, for an account A on one site and an account B on another, whether the two are owned by the same person.
- We tried three classifiers: Decision trees, Naive Bayes, and Support Vector Machines.
- We defined features that measure the similarity of textual content, temporal and geospatial activity patterns, and social connections across two accounts.

4. Similarity features

- Our similarity features are based on comparing vectors of activity patterns of different types. In particular, we use:
  - **Jaccard similarity:**

$$F_1(u, v) = \frac{\sum_i \min(u_i, v_i)}{\sum_i \max(u_i, v_i)}$$

  - **Hamming similarity:**

$$F_2(u, v) = \sum_i 1(u_i = v_i)$$

  - **Cosine similarity:**

$$F_3(u, v) = \frac{u \cdot v}{\|u\| \, \|v\|}$$

- We also used variations, including normalizing by bin popularity (as in TF-IDF weighting), and a statistic measuring probability of the similarities occurring by random chance.

- For **textual similarity**, we build a vector for each account consisting of a histogram over all possible words.
  - I.e. bags-of-words in classic vector space model.
  - Use all words in tweets, and all tags in Flickr photos.

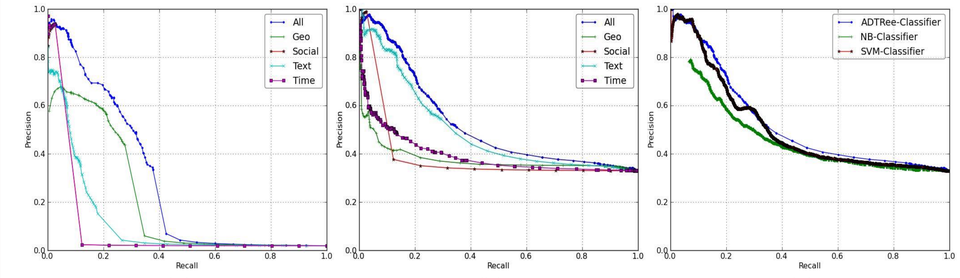
- For **temporal similarity**, we discretize time into bins and count the number of activities (e.g. tweets or photos) in each bin, to produce a high-dimensional vector.
  - Tried two discretizations: 1 bin per day, and 1 bin per hour.

- For **geospatial similarity**, build vectors over place names.
  - Map GPS coordinates and user-reported location strings to canonical town names (via the GeoNames database).
  - Also build geo-temporal vectors, i.e. in Cartesian product of geospatial and temporal spaces.

- For **social connections similarity**, simply examine the two sets of social connections, and count number of usernames in common.

5. Results

- We partitioned data into training and test sets, and then trained each classifier using the ground truth correspondences.
- Then hid the ground truth labels and tested ability of classifiers to retrieve pairs of accounts owned by the same user, in terms of precision and recall:



(a): Performance for different features with the decision tree classifier.
(b): Performance for different features with decision trees, when negative exemplars are chosen only amongst users in the same hometown.
(c): Performance of different classifiers using all features.

6. Summary and future work

- Examined whether relatively weak features can be used to identify accounts belonging to the same person across different sites.
- On our dataset from Flickr and Twitter, we find that **about 10% of corresponding accounts can be successfully de-anonymized at a precision of about 90%**.
- Future work includes larger datasets over longer periods, and more sophisticated similarity features.

7. Acknowledgments

We thank **Michael Conover** for collecting the Twitter data and for many helpful discussions, **Kris Hauser** for suggesting the machine learning approach, and **Kiran Kumar** for the initial exploratory work on this topic.