Brave New Task: User Account Matching

Pisa – October 5, 2012

C. Hauff, G. Friedland
TU Delft & ICSI
Users on the Social Web

Not just one account, but many accounts.
Users on the Social Web

“Cooperative” users: publicly provide their respective accounts.

Not just one account, but many accounts.
User Account Matching

Can we identify the same user in another social Web stream universe?
User Account Matching

- Different scenarios

1 vs. 1

Our task setup.

1 vs. k

additional evidence
Why?

- **Benevolent uses**
  - Enriched user models
  - Improved personalization effectiveness
  - To make users happier 😊

- **Malicious uses**
  - Password recovery (self-service password reset) on a large scale
  - Discover “offline” information based on enriched profiles (e.g. phone numbers)

---

**Example Recovery Questions**

- What is your favorite team?
- What is your favorite movie?
- What is your favorite TV program?
- What is your least favorite nickname?
- What is your favorite sport?
- Who was your childhood hero?
- What was the first concert you attended?
- What time of the day were you born?
- What was your dream job as a child?
- What is the middle name of your oldest child?

Source: [http://goodsecurityquestions.com](http://goodsecurityquestions.com)
Existing work with a strong text-based bias

- Most previous work based *directly* on the profile information
Existing work profile-information based

- Zafarani et al., 2009
  - Similarity of *user names* on different platforms
  - Automatic matching ground truth: BlogCatalog (cooperative users)

- Abel et al., 2010
  - Investigated the amount of user profile aggregation possible with cross-community linking (cross-links retrieved from the Social Graph API)

Source: Abel et al., Interweaving Public User Profiles on the Web, UMAP 2010
Existing work beyond profiles

• Narayanan et al., 2009
  • Rely on the graph structure of social networks to de-anonymize the graph (no use of profile or content information)

• Iofciu et al., 2011
  • Used tags (StumbleUpon, Flickr, Delicious) of images and bookmarks to identify matching accounts
  • Ground truth based on the Social Graph API
  • Content-based matching (compared to user name matching) is a much more difficult task
    • Starting point for our work
Our task (1 vs. 1)

Given a Flickr account ...

determine the corresponding Twitter account from a large set of potential streams.
Our task (1 vs. 1)

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 matches?

Given a Flickr account ...

determine the corresponding Twitter account from a large set of potential streams.
Data Set: The Basics

• 50,000 semi-random users selected on Twitter and followed for three months (04/2012-06/2012)
  • ~18,000 tweeted at least once in that time period

• Manually checked potential matching Flickr accounts
  • Potential matches: (i) tweets containing flickr.com,
    (ii) existing Flickr account with the same user name

<table>
<thead>
<tr>
<th>Number of account pairs</th>
<th>233</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of tweets in account pairs</td>
<td>214,664</td>
</tr>
<tr>
<td>Number of images in account pairs</td>
<td>245,320</td>
</tr>
</tbody>
</table>

| Number of Twitter accounts $N$ | 18,372 |
| Number of tweets in $T$ | 2,795,388 |
Data Set: User Distribution

- 200 photos (Flickr limit)
- more tweets than images
- more images than tweets
Data Set: The Temporal Dimension

119 account pairs with overlapping time stamps

No information available
Baseline

• Treat all tweets of a user as **document**
  • **Corpus** of Twitter user documents

• Treat all textual information from a user’s Flickr stream as a (very long) **query**

• **Rank the documents with respect to the query**
  (i.e. rank the Twitter accounts)

*This is a standard ad hoc retrieval problem: we used Okapi.*
Baseline Results

Account matching based on content is hard.

The larger the number of Flickr images, the better the matching.

RR = \frac{1}{\text{rank}(\text{matching Account})}

<table>
<thead>
<tr>
<th></th>
<th>τ #tweets</th>
<th>τ #images</th>
<th>τ #tweets + #images</th>
</tr>
</thead>
<tbody>
<tr>
<td>MRR</td>
<td>0.168</td>
<td>0.033</td>
<td>0.200</td>
</tr>
</tbody>
</table>
Baseline Results: Taking a Closer Look

- Distribution of the 233 RR values

- Influence of time overlap in MRR

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>119 account pairs with overlap</td>
<td>0.2134</td>
</tr>
<tr>
<td>114 account pairs without overlap</td>
<td>0.1253</td>
</tr>
</tbody>
</table>

Task is either very easy or very difficult. Less than 20% of ‘queries’ with non-0/1 RR.

Time overlap in streams makes the task easier!
Challenges

① Social networks have (strong) **data gathering restrictions** in place – requires long term setup
  • Twitter: complete user history not available
  • Flickr: max. 200 photos for users without “pro” accounts

② Users use different social networks **at different time periods** – makes the matching even more difficult

③ **Automatic** ground truth generation is **error-prone**: self-reported links can be arbitrary, link to friends, etc.
  • Crowd-sourcing may be an option

④ Many encountered matches are not from private individuals, but belong to **organizations or businesses**
Challenges

1. Social networks have strong restrictions in place – requires long term setup
   - Twitter: complete user history not available
   - Flickr: max. 200 photos for users without “pro” accounts

2. Users use different social networks at different time periods – makes the matching even more difficult

3. Automatic ground truth generation is error-prone: self-reported links can be arbitrary, link to friends, etc.
   - Crowd-sourcing may be an option

4. Many encountered matches are not from private individuals, but belong to organizations or businesses
Thank you!

All suggestions are welcome!