Problems and methods for attribute detection of social network users

Anton Korshunov

Institute for System Programming of Russian Academy of Sciences

RCDL-2013
1 Network Level: User Community Detection

2 User Level: Demographic Attribute Detection

3 Inter-network Level: User Identity Resolution
Contents

1 Network Level: User Community Detection

2 User Level: Demographic Attribute Detection

3 Inter-network Level: User Identity Resolution
Communities: Definition

Functional definition of communities

Communities serve as organizing principles of nodes in social networks and are created on shared affiliation, role, activity, social circle, interest or function.

Cover

Cover of a social graph is a set of communities such that each node is assigned to at least one community.
Facebook Friendship Graph: Global Communities
Communities: Structural Properties

Structural properties of communities

- **Separability**: good communities are well-separated from the rest of the network
- **Density**: good communities are well connected
- **Cohesiveness**: it should be relatively hard to split a good community
Applications

Traffic optimization
Traffic inside communities is more intensive, so it makes sense to place all nodes comprising large communities onto the same data node/warehouse.

Link and attribute prediction
Thanks to the homophily principle of community organization, users inside communities tend to have similar attribute values and increased probability of establishing new links.

Graph closeness
Estimating how close are nodes in the social graph is possible by comparing their community memberships.

Spam detection
It is possible to not only detect single spammers by analyzing their content, but to detect spam networks by analyzing links.

Recommender systems
Enhancing social recommendation systems with a-priori known groupings of users.
Task Definition

Input
- social graph
- algorithm parameters

Output
Found cover of global communities (user-community assignments)
Requirements

- **Ability to discover overlapping community structure**
  People tend to split their social activities into different circles

- **Support for directed edges**
  Directed edges (parasocial relationships) are common in content networks

- **Support for weighted edges**
  Edge weights could be used to add apriori knowledge about similarity of users

- **High accuracy**
  The algorithm must prove its applicability to real and synthetic graphs

- **Efficiency**
  The algorithm must have low computational complexity

- **Distributed version**
  The algorithm must be runnable in cloud environment (e.g., Amazon EC2)
**Speaker-listener Label Propagation Algorithm (SLPA)**

1. The memory of each node is initialized with a unique community label.
2. The following steps are repeated until the maximum iteration $T$ is reached:
   a. One node is selected as a listener.
   b. Each neighbor of the selected node randomly selects a label with probability proportional to the occurrence frequency of this label in its memory and sends the selected label to the listener.
   c. The listener adds the most popular label received to its memory.
3. The post-processing based on the labels in the memories and the threshold $r$ is applied to output the communities.
Approach: Speaker-listener Label Propagation Algorithm

**Advantages**

1. Able to uncover overlapping/disjoint global/local community structure
2. Supports directed edges and edge weights
3. High accuracy
4. $O(T \cdot |E|)$ complexity ($|E|$ – number of edges in the graph)
5. Easy distributable in a natural way
Approach: Initialization Using Maximum Cliques

Idea

- Extract maximum cliques with at least $k$ nodes
- Assign the same label to all nodes within a single clique
- Communities tend to organize themselves around cliques

Conrad Lee et al. 2010
Detecting Highly Overlapping Community Structure by Greedy Clique Expansion
Approach: Specific Interaction Rules for Local Communities

Idea

**Local community** - a community of a user’s contacts

- Find local communities for each node
- Listener accepts 1 most frequent label from each local community at each iteration
- Resulting global communities *inherit* the structure of local communities

Local Community Detection

1. Extract ego-network (1.5-neighbourhood) of each user
2. Apply SLPA to the user’s ego-network
Sample graph by LFR benchmark: $N = 120, O_n = 10, O_m = 6$

Normalized Mutual Information (NMI) of covers $X$ and $Y$

$$NMI(X : Y) = 1 - \frac{1}{2}[H(X|Y)_{norm} + H(Y|X)_{norm}]$$
Accuracy Evaluation

Undirected non-weighted graphs by LFR benchmark

$N=2000; \ Om=4; \ threshold=0.05; \ 20 \ iterations$

![Graph showing NMI vs. On/N for different methods]

- plain vanilla SLPA
- rough cores
- Cliques 5
- egommunities
- RC + egommunities
- Cliques 5 + egommunities
Performance Evaluation: Scalability by Graph Size

Spark.Bagel implementation @ Amazon EC2

- \textit{threadsCount} = 80
Spark.Bagel implementation @ Amazon EC2

- $|V| = 1M$
1 Network Level: User Community Detection

2 User Level: Demographic Attribute Detection

3 Inter-network Level: User Identity Resolution
## Demographic Attributes

### Categorical
- gender
- relationship status
- social status
- education level
- political views
- religious views
- ...

### Integral
- age
- income
- ...

Attribute Values of Twitter Users

Gender Distribution on Twitter

- Male: 47%
- Female: 53%

Source: www.beevolve.com
Self Disclosed Age Distribution on Twitter

Source: www.beevolve.com
Problems

- **missing attributes**
  - Name: Julia Stevens
  - Age: [empty]
  - Gender: female
  - Relationship: [empty]
  - Location: France

- **(un)intended mistakes**
  - Name: Rob Fee
  - Age: 666
  - Gender: female
  - Relationship: single
  - Location: U.S.

- **stolen/false identities**
  - Name: Maria Zotova
  - Age: 24
  - Gender: female
  - Relationship: married
  - Location: Moscow
**Task Definition**

**Input**
- user tweets
- user profile
- algorithm parameters

**Output**
Values of predicted attributes

<table>
<thead>
<tr>
<th>Gender</th>
<th>Age</th>
<th>Relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>MALE</td>
<td>&lt;20</td>
<td>SINGLE</td>
</tr>
</tbody>
</table>
Issues

- Informal chatter style
- Lots of mycrosyntax, slang, abbreviations and spelling mistakes
- Limited message length
- Manual labeling of training set is time-consuming
- High dynamicity of Twitter language → periodical retraining is required
- Lots of citations (retweets) → lack of original text
Approach

1. **Building training sets**
   - **languages**: EN, RU, DE, FR, IT, ES, PT, KO
   - **attributes**: gender, age, relationship status, political and religious views

2. **Preprocessing**
   - removing retweets
   - filtering by language

3. **Binary feature extraction**
   - **sources**: raw tweet texts and user profiles
   - **features**: [1..7]-grams over cased/uncased characters and tokens

4. **Feature selection**
   - Conditional Mutual Information

5. **Model learning**
   - Online Passive-aggressive Algorithm

6. **Classification**
Advantages

- Automatic compilation
- Support of multiple user attributes through Facebook
- Multilinguality
Original tweet:

It's times like this that I wish I had a boyfriend to cuddle up to and cry on.

More from this user:

Oh great now I need gas too.
I wannaa gooooo fishingggg
My kind of your kind of it's this kind of night, we dance in the dark and your lips land on mine.
My Kinda Night just came on the radio.

☐ IF YOU HAVE A TRAILER THAT CAN BE USED FOR A HOCO FLOAT

TWEET/TEXT ME OR CHAD. StuCo is in desperate need!

Brooks is a lifesaver
The jeep leaks.
A random stranger propelled by the will of God can be the person that blesses you the most. God is so good. Forever in awe of His glory.
My hoco group's shirt is better than yours.

Gender: female
Age: middle
Relationship status: single
Political views: democrat
Religion: christian
Language: English
Country: unknown
Original tweet:

Haven't been to sleep yet n my husband already left for work -_____

More from this user:

Now I have Andra in bed with me, maybe I'll fall asleep soon.

Enjoyed coloring n talking with my cousin @AmandaNoelle73 though, thanks for keeping me company ♥

@AmandaNoelle73 haha ok

@AmandaNoelle73 not really lol table then?haha

Need something to drink...with lots of ice #parched

@AmandaNoelle73 I have my light on haha they're both knocked out so they dont notice..wanna come to the room or meet you at the table?lol

@AmandaNoelle73 it's hard to..Nick did the other night though..right now he's just wiggling if he moves more

I'll find you lol

Interesting evening...honestly did much better than I expected which is good

Coloring #boredaf

---

Gender: female

Hide:

'♥' [char_name] 0.144310863805

'nn' [char_screen_name] 0.126487818889

'na' [char_screen_name] 0.110117264848

'y husba' [char] 0.104609059447

'♥' [char] 0.0935898338195

'my h' [char_uncased] 0.073603183063

'ia' [char_screen_name] 0.0710998703084

've my' [char_uncased] 0.0550267494366

'nterm' [char_uncased] 0.0347557671199

'_' [char] 0.0309442704324

Age: middle
## Accuracy Evaluation

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Users</th>
<th>Tweets</th>
<th>Accuracy</th>
<th>Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>age (birthdate)</td>
<td>1180</td>
<td>56640</td>
<td>69.1%</td>
<td>65.0%</td>
</tr>
<tr>
<td>age (+year of graduation)</td>
<td>3755</td>
<td>180240</td>
<td>71.4%</td>
<td>63.3%</td>
</tr>
<tr>
<td>gender (profile)</td>
<td>17050</td>
<td>818400</td>
<td>83.3%</td>
<td>50.0%</td>
</tr>
<tr>
<td>gender (+dictionary)</td>
<td>70734</td>
<td>3395424</td>
<td>89.2%</td>
<td>50.0%</td>
</tr>
<tr>
<td>relationship status</td>
<td>1901</td>
<td>202175</td>
<td>89.0%</td>
<td>%</td>
</tr>
<tr>
<td>political views</td>
<td>662</td>
<td>31776</td>
<td>73.7%</td>
<td>53.8%</td>
</tr>
<tr>
<td>religious views</td>
<td>1491</td>
<td>71568</td>
<td>88.0%</td>
<td>76.5%</td>
</tr>
</tbody>
</table>

- **English users**
- 48 original (non-retweet) tweets for each user
- baseline corresponds to classification into the most common class
Accuracy Evaluation: Impact of Non-confidence
Contents

1 Network Level: User Community Detection

2 User Level: Demographic Attribute Detection

3 Inter-network Level: User Identity Resolution
Overlap of Social Network Populations

SNS Usage Overlap

Facebook
MySpace
Twitter
LinkedIn

Source: Anderson Analytics 2009
Aligning & Merging Social Graphs

Benefits

- Allow cross-platform information exchange and usage
- Enrich existing profiles with data from other networks
- Cold-start problem solving
Contact Lists Merging
Task Definition

**Input**

Two different ego-networks \(< A, B >\) of a single user:

- Profile attributes (name, birthday, home town, ...)
- Social links (friendship, subscription, ...)

**Output**

All profile pairs \((v, u)\mid v \in A, u \in B\) that belong to the same real person
Main idea
If \( v \) and \( u \) are connected in graph \( A \) than their matches \( \mu(v) \) and \( \mu(u) \) should be as similar as possible in graph \( B \)

Criteria for choosing projections
- How similar is \( v \) to its possible projection based on similarity of profile fields?
- How many contacts a possible projection shares with projections of neighbours of \( v \)?
Steps

1. Build *Conditional Random Fields* model from Twitter and Facebook graphs
2. Estimate *anchor nodes* (a-priori known projections)
3. Compute *edge energies*
   - profiles: string similarity of fields
   - graph: weighted Dice measure
4. Find the optimal configuration of matching nodes
5. Filter the results by pruning unwanted matches

Sergey Bartunov, Anton Korshunov et al
Joint Link-Attribute User Identity Resolution in Online Social Networks
*The 6th SNA-KDD Workshop August 2012, Beijing, China*
Accuracy Evaluation

Results

<table>
<thead>
<tr>
<th>algorithm</th>
<th>R</th>
<th>P</th>
<th>F₁</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline 1 (weighted sum)</td>
<td>0.45</td>
<td>0.94</td>
<td>0.61</td>
</tr>
<tr>
<td>Baseline 2 (probability distance)</td>
<td>0.51</td>
<td>1.0</td>
<td>0.69</td>
</tr>
<tr>
<td>Joint Link-Attribute model</td>
<td>0.8</td>
<td>1.0</td>
<td>0.89</td>
</tr>
</tbody>
</table>

Dataset

<table>
<thead>
<tr>
<th></th>
<th>Twitter</th>
<th>Facebook</th>
</tr>
</thead>
<tbody>
<tr>
<td># of seeds</td>
<td></td>
<td>16</td>
</tr>
<tr>
<td># of profiles</td>
<td>398</td>
<td>977</td>
</tr>
<tr>
<td># of connections</td>
<td>1728</td>
<td>10256</td>
</tr>
<tr>
<td># of matches</td>
<td></td>
<td>141</td>
</tr>
<tr>
<td># anchor nodes</td>
<td></td>
<td>71</td>
</tr>
</tbody>
</table>
Baseline

Optimal matching as an assignment problem

Similarity functions

1. weighted sum of profile similarity vector \( V(v, \mu(v)) \)
2. \( 1 - \text{profile-distance}(v, \mu(v)) \)
QUESTIONS ?