Aggregating Content and Network Information to Curate Twitter User Lists

Derek Greene
Gavin Sheridan

Barry Smyth
Pádraig Cunningham

RecSys-12: Recommender Systems & The Social Web
Task: User List Curation

- Twitter allows the grouping of users into topical **user lists**.
- **Storyful** maintains lists of users for news stories to monitor breaking news related to that story.
Supporting User List Curation

- **Idea:** Use data analysis to identify important users that form the “community” around a news story on Twitter.

- **Goal:** Recommend new users to expand an embryonic seed list, helping to find valuable content relating to the story.
Curation System Overview

Seed Set

Bootstrap

Candidate List

Recommender

Ranked List

Curator

Candidate List

Core List

Updater

Phase 1

Phase 2
Recommendation Overview

General Recommender

- Use embryonic seed list as initial training data.
- Compute training data centroid.
- Measure cosine similarity between centroid and vectors representing candidate non-seed users.

→ Produce ranked list of top $K$ non-seed users to present to a human curator.

Which criterion should we use for recommendation?

→ Propose variety of vectorised approaches, using criteria based on Twitter network and content analysis.
Content-Based Criteria

• **Tweet profiles:** Create a term vector for each user, containing the aggregation of their 50/100/200 most recent tweets.

<table>
<thead>
<tr>
<th>User</th>
<th>president</th>
<th>health</th>
<th>care</th>
<th>reform</th>
<th>gender</th>
<th>costs</th>
</tr>
</thead>
<tbody>
<tr>
<td>@BarackObama</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Content-Based Criteria

- **List names/descriptions:** Create a term vector for each user, containing the aggregation of names and/or descriptions of user lists to which they have been assigned.

<table>
<thead>
<tr>
<th>User</th>
<th>elected</th>
<th>politics</th>
<th>purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>@SarahPalinUSA</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>User</th>
<th>politics</th>
<th>news</th>
<th>useless</th>
</tr>
</thead>
<tbody>
<tr>
<td>@SarahPalinUSA</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
# Network-Based Criteria

- **Followed-by Profiles:** Users are similar if they are "co-followed" by the same set of users. Represent each user as a binary follower profile vector.

  *User X is followed by...*

<table>
<thead>
<tr>
<th>User X</th>
<th>@TeamGB</th>
<th>@bradwiggins</th>
<th>@Mo_Farah</th>
<th>@britishswimming</th>
</tr>
</thead>
<tbody>
<tr>
<td>@MichaelPhelps</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>@chrishoy</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>@TomDaley1994</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

- **Mentioned-by Profiles:** Users are similar if they are mentioned in tweets by the same users (i.e. "co-mentioned").

- **Retweeted-by Profiles:** Users are similar if their posts are retweeted by the same users (i.e. "co-retweeted").
Network-Based Criteria

- Co-Listed Information:
  - Other media outlets and individuals will also be simultaneously curating user lists on topics in the wider Twitosphere.
  - Would like to “crowdsource” these efforts to support list curation.

Users @BrianODriscoll and @KearneyRob are co-listed on both lists.
Experimental Evaluation

• Collected 10 datasets around "Super Tuesday" GOP nominations in March 2012, with annotated seed users from Storyful.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Users</th>
<th>Core Users</th>
<th>Tweets</th>
<th>Friends</th>
<th>Followers</th>
<th>Listed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alaska</td>
<td>948</td>
<td>41</td>
<td>185</td>
<td>208</td>
<td>269</td>
<td>89</td>
</tr>
<tr>
<td>Georgia</td>
<td>966</td>
<td>34</td>
<td>211</td>
<td>235</td>
<td>295</td>
<td>126</td>
</tr>
<tr>
<td>Idaho</td>
<td>743</td>
<td>20</td>
<td>186</td>
<td>264</td>
<td>273</td>
<td>47</td>
</tr>
<tr>
<td>Massachusetts</td>
<td>821</td>
<td>24</td>
<td>209</td>
<td>244</td>
<td>293</td>
<td>122</td>
</tr>
<tr>
<td>North Dakota</td>
<td>363</td>
<td>26</td>
<td>203</td>
<td>147</td>
<td>192</td>
<td>93</td>
</tr>
<tr>
<td>Ohio</td>
<td>1051</td>
<td>97</td>
<td>178</td>
<td>171</td>
<td>207</td>
<td>115</td>
</tr>
<tr>
<td>Oklahoma</td>
<td>693</td>
<td>32</td>
<td>205</td>
<td>178</td>
<td>211</td>
<td>109</td>
</tr>
<tr>
<td>Tennessee</td>
<td>979</td>
<td>48</td>
<td>199</td>
<td>170</td>
<td>204</td>
<td>112</td>
</tr>
<tr>
<td>Vermont</td>
<td>864</td>
<td>36</td>
<td>182</td>
<td>169</td>
<td>190</td>
<td>66</td>
</tr>
<tr>
<td>Virginia</td>
<td>877</td>
<td>46</td>
<td>200</td>
<td>160</td>
<td>199</td>
<td>115</td>
</tr>
</tbody>
</table>
Experimental Setup

- Ran 250 X \( k \)-fold cross validation experiments per dataset:
  - Held out a proportion of the seed set as test data.
  - Used the remaining seed users as training data.
  - Ranked users in test data using each criterion.
  - Calculated \textit{precision} and \textit{recall} relative to the test data.
Comparison of Criteria

- Looked at diversity among the recommendations from the different criteria...

→ We see several distinct signals present across the different views.

→ No single criterion performs consistently well on all datasets.
• **SVD Aggregation**: Combine rankings generated on top 5 individual criteria into a single matrix, apply Singular Value Decomposition, then rank values in first singular vector.

**Comparison of SVD Aggregation V Top 5 Individual Criteria**
User List Curation System

**User Lists**
- Create New List...
- Import List...
- Manage Blacklists
- **>> afghanistan**
- **>> africa**
- **>> alabama**
- **>> albania**
- **>> algeria**
- **>> andalusia**
- **>> angola**
- **>> argentina**
- **>> arizona**
- **>> armenia**
- **>> art**
- **>> asia-pacific**
- **>> austria**
- **>> azerbaijan**
- **>> bahrain**

**Curation System Home**

**Recent Activity:**
- Last bootstrapped: [greece] [portugal] [uk]
- Last recommended: [greece] [france] [it-security]
- Last updated: [greece] [france] [it-security]
- Last modified: [master-list]

**Pending Instructions:**

<table>
<thead>
<tr>
<th>Project</th>
<th>Process Type</th>
<th>Requested At</th>
<th>Scheduled For</th>
</tr>
</thead>
<tbody>
<tr>
<td>it-security</td>
<td>CONFIGURE</td>
<td>2012-06-22 12:08</td>
<td>2012-06-22 12:08</td>
</tr>
<tr>
<td>angola</td>
<td>UPDATE</td>
<td>2012-06-22 01:04</td>
<td>2012-06-22 12:16</td>
</tr>
<tr>
<td>afghanistan</td>
<td>RECOMMEND</td>
<td>2012-06-22 11:51</td>
<td>2012-06-22 12:24</td>
</tr>
<tr>
<td>portugal</td>
<td>UPDATE</td>
<td>2012-06-22 01:22</td>
<td>2012-06-22 12:35</td>
</tr>
<tr>
<td>france</td>
<td>UPDATE</td>
<td>2012-06-22 02:52</td>
<td>2012-06-22 13:22</td>
</tr>
</tbody>
</table>

**Active Processes:**

<table>
<thead>
<tr>
<th>Project</th>
<th>Process Type</th>
<th>Started At</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>armenia</td>
<td>BOOTSTRAP</td>
<td>2012-06-22 11:59</td>
<td>RUNNING</td>
</tr>
<tr>
<td>kosovo</td>
<td>BOOTSTRAP</td>
<td>2012-06-22 11:58</td>
<td>RUNNING</td>
</tr>
<tr>
<td>uk</td>
<td>UPDATE</td>
<td>2012-06-22 11:48</td>
<td>RUNNING</td>
</tr>
</tbody>
</table>
Conclusions

• **Summary:**
  - Proposed variety of criteria for user list building.
  - Performed a comprehensive comparison of individual criteria, demonstrating weaknesses of both content and network-based strategies.
  - Demonstrated that more accurate results can be achieved using SVD aggregation.

• **Future Work:**
  - Stratification of networks to target recommendations for niche communities - avoiding the "filter bubble".
  - Support for curation across multiple social networks.
Any Questions?

derek.greene@ucd.ie
@derekgreene
http://cliquecluster.org