Discussion Graphs: Putting Social Media Analysis into Context

Emre Kıciman, Scott Counts, Michael Gamon
Microsoft Research

Munmun De Choudhury
School of Interactive Computing
Georgia Tech

Bo Thiesson
Dept. of Computer Science
Aalborg University
Social media is a rich source of information

• Relationships between drugs, symptoms and side-effects [Paul and Dredze 2011]

• Disease transmission based on co-visited locations [Sadilek et al. 2012]

• User behaviors and hashtags in context of Mexican drug war [Monroy-Hernandez et al. 2013]

• Relationships between locations based on co-visits by users [Cranshaw et al. 2012]
Context is critical

• Context == where did the discussion come from?
  • Temporal, spatial, topical, demographic and other

• Critical for interpreting results
• Conditioning on context appropriately can change results

• Not common in large-scale quantitative SM analyses
The rest of this talk...

- Goal: Simplify analysis of social media data in context
  - Simplify extraction of features
  - Simplify contextual conditioning and tracking
  - Focus on co-occurrence analyses

- Discussion graphs: data model and analysis tool
- Case studies
Co-occurrence analysis

• Two things are related if they co-occur together

• Basic analysis technique. Often a building block.
Co-occurrence examples

In a message:
“I’m eating a donut and coffee”
“I love dipping donuts into milk”
Co-occurrence examples

Related by co-visits

Bob checked in at “Santiago Airport”
Bob checked in at “Sheraton Hotel”
Co-occurrence examples

Activities co-occurring within the same moods:
  “relaxing” ~ “listening to music”

People related by co-occurring within the same locations

And others...
What’s the trouble?

• Conceptually, it is straightforward.

• But many practical challenges
  • Building and sharing feature extractors
  • Long, messy scripts
  • Counting weighted features
  • “Debugging” (e.g., sampling supporting tweets)

• Result $\Rightarrow$ slows down iterations and depth of analysis
Discussion graphs

Formalize co-occurrence analysis using hyper-graphs as a data model

Small number of operations capture co-occurrence analyses

1. **EXTRACT**: What features should be extracted from social messages?
2. **RELATE**: What defines a relationship?
3. **PROJECT**: What is the domain of relationships to extract?

Result provides a succinct representation of arbitrary analyses

- Easy to write and modify
- Analysis can be automatically augmented with “best practices”
“I had fun hiking Tiger Mountain last weekend” – Alice said on Monday, at 10am
Relate

• Declare a new relationship through domain $R$

• All hyper-edges connected to a node in $R$ become a single hyper-edge
Name: Alice
Location: Tiger Mountain
Gender: Female
Mood: Happy
Activity: Hiking
Post Time: Mon 10am

Name: Bob
Gender: Male
Post Time: Fri 3pm
Activity Time: {Sat-Sun}
Name: Alice
Gender: Female
Mood: Happy
Post Time: Mon 10am
Activity: Hiking

Name: Bob
Gender: Male
Post Time: Fri 3pm
Activity Time: {Sat-Sun}
Projection

• Often want to limit structural analyses to some small number of domains

• Restrict graph structure to only include nodes in target domains

• Aggregate all other domains as contextual statistics; associate with edges in new projected graph
Name: Alice
Location: Tiger Mountain
Gender: Female
Mood: Happy
Post Time: Mon 10am
Activity: Hiking

Name: Bob
Gender: Male
Post Time: Fri 3pm
Activity Time: {Sat-Sun}
• Analyze key relationships by projecting graph to lower dimensions
• Aggregate statistical distribution of other dimensions as context for remaining edges
• Here, we analyze relationship between locations and activities. Information about gender, time-of-day, etc. is a summary context for edge
Discussion Graph Tool

• Implemented this model in a simple scripting tool

• Includes several common feature extractors
  • Tokens, time features, phrase extractors, Wikipedia entity linking
  • Sentiment/mood extractor
  • Author statistics, gender, hometown, county

• Incorporates/enables best practices
  • Keeps random sample of supporting raw data for each relationship
  • Tracks provenance of outputs
  • Easy to mix-and-match features; iterate on analyses
Example script: Activity-Location relations

LOAD  Twitter(startdate:“9/15/12”,
   enddate:“10/15/12”);

EXTRACT
   PRIMARY PhraseMatch(match:“locationlist.txt”,
      domain:“location”),
   PRIMARY PhraseMatch(match:“activitylist.txt”,
      domain:“activity”),
   Mood(), Gender(), County(), Time();

PROJECT TO  location, activity;
OUTPUT TO  “location-activity.graph”;
Result sample: Activity-Location relations

- vacation

<table>
<thead>
<tr>
<th>Location</th>
<th>Number of Cooccurrences</th>
<th>Association Strength (PMI)</th>
<th>Sentiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hawaii</td>
<td>687</td>
<td>3.63</td>
<td></td>
</tr>
<tr>
<td>Martha’s Vineyard</td>
<td>419</td>
<td>4.10</td>
<td></td>
</tr>
<tr>
<td>Florida</td>
<td>386</td>
<td>4.24</td>
<td></td>
</tr>
<tr>
<td>Miami</td>
<td>252</td>
<td>4.67</td>
<td></td>
</tr>
</tbody>
</table>
Example script: Political issues

**LOAD**  Twitter(startdate: "9/15/12",  
enenddate: "10/15/12");

**EXTRACT**

- PRIMARY PhraseMatch(match: "politicianlist.txt",  
  domain: "politician"),
- PhraseMatch(match: "issueslist.txt",  
  domain: "issue"),
- Mood(), Gender(), County(), Time();

**PROJECT TO**  politician, absoluteday;

**OUTPUT TO**  "politicianPerDay.graph";
Result: Political issues
2 Brief Case Studies using Context

1. Context helping interpret higher-level graph structures

2. Higher-level analyses vary based on original context
#1: Summary Context for Graph Structures

- Atlas
- Central
- New York City
- Prometheus
- Manhattan
- New York
- Rockefeller Center
- Central Park
- Music Hall
- Empire State Building
- 49th Street
- Midtown

- World Trade Center
- Empire State Building
- Cityscape
- Park Avenue
- Manhattan
- Trump World Trade Center
- Midtown
#1: Summary Context for Graph Structures

<table>
<thead>
<tr>
<th></th>
<th>New York Tourist</th>
<th>Midtown Worker</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>49%</td>
<td>63%</td>
</tr>
<tr>
<td>Female</td>
<td>33%</td>
<td>23%</td>
</tr>
<tr>
<td>Metroarea</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NYC</td>
<td>33%</td>
<td>54%</td>
</tr>
<tr>
<td>Other</td>
<td>67%</td>
<td>46%</td>
</tr>
<tr>
<td>Mood</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Joviality</td>
<td>56%</td>
<td>49%</td>
</tr>
<tr>
<td>Fear</td>
<td>14%</td>
<td>13%</td>
</tr>
<tr>
<td>Sadness</td>
<td>11%</td>
<td>15%</td>
</tr>
<tr>
<td>Guilt</td>
<td>8%</td>
<td>6%</td>
</tr>
<tr>
<td>Fatigue</td>
<td>3%</td>
<td>6%</td>
</tr>
<tr>
<td>Serenity</td>
<td>3%</td>
<td>4%</td>
</tr>
<tr>
<td>Hostility</td>
<td>2%</td>
<td>4%</td>
</tr>
</tbody>
</table>
#2: Higher-level analyses: Neighborhoods

- Infer social distance between locations based on co-visits
- Mix with geographic distance and cluster
  (Cranshaw et al. ICWSM 2012)

- Same analysis conditioned on temporal and demographic factors
- Data: 2.3 M geo-located tweets in NYC, Jan. 1 to Mar. 31, 2013
Weekday / Weekend

Weekday

Weekend
By Gender

Male

Female
Conclusions

• Discussion graphs simplify co-occurrence analyses
  • Formal definition and constrained, domain-specific language
  • Succinct representation of a common class of analyses
  • Tooling automates common tasks/best practices

• Discussion graphs make it easy to capture and condition on context
  • Context helps interpret higher-level results
  • Conditioning on context can dramatically change results

• For more details, see upcoming paper in ICWSM-14.


• Questions?  Contact emrek@Microsoft.com