Rhetorical Use Case

Jeff Reminga and Neal Altman

The CASOS Center
COS Program, School of Computer Science, Carnegie Mellon
Summer Institute 2018

Overview

- Demonstration of a workflow in ORA for text analysis
- You have a set of documents, tweets, email, blogs etc.
- You are interested in understanding what is being said and identifying the messages with the greatest influence and potential reach. You want to characterize what is being discussed.
Input Data

- Data to be analyzed has the following nodesets:
  - **Author**: This is the person, organization, group, or both that created the text. Examples: the Tweeter who sent a tweet, the journalist who wrote the news article, or the sender of the email.
  - **Document**: This is the way of identifying the text. This might be the twitter id. This might be the title of an article or book.
  - **Concept**: This is the unit into which the text is broken down. This might be a single word e.g. boy, or it might be a phrase – e.g., United States of America. Hashtags are often used for Twitter.

Example data: Automap

- Meta-networks output from Automap contain:
  - Concept nodeset
  - Semantic network (which is Concept x Concept)
Step 1: Get Data

- Any set of texts, e.g., news articles or blogs. This might be a set of media-posts, e.g. tweets or emails. If you are going to use AutoMap – remove all images and have one “document” per file. If importing tweets you can import the JSON directly into ORA.

Example data: Twitter

- Twitter data imported using ORA contains:
  - Concepts: Hashtags or Words
  - Authors: Agents
  - Document: Tweet
  - Semantic network: Hashtag x Hashtag – co-occurrence, Word x Word – co-occurrence
  - Concept x Author: Hashtag x Agent
  - Author x Document: Agent x Tweet
Example data: Twitter

- Twitter data imported using ORA contains:
  - Concepts: Hashtags or Words
  - Authors: Agents
  - Document: Tweet
  - Semantic network: Hashtag x Hashtag – co-occurrence, Word x Word – co-occurrence
  - Concept x Author: Hashtag x Agent
  - Author x Document: Agent x Tweet

Twitter importer

- Data Import Wizard imports Twitter JSON
Twitter importer

- Automatically creates the data needed for rhetorical analysis
- Using hashtags or using words found in tweet message

Step 2: Create Usage Network

- Note in AutoMap and Netmapper, and in the ORA twitter importer, Step 2, 3 and 4 are done at the same time

- Create usage network. This network says what concept is used in which document or by which author (the “writer” or “speaker”). Convert this data to a document by concept, or author by concept matrix. The link is weighted and represents the number of times that concept is used in that text or by that author.
Step 3: Create Semantic Network

- This is a network of concepts by concepts such that the link between the concepts is the number of documents or authors that use both concepts.
- Automap: outputs automatically Semantic Network
- ORA Twitter importer will create:
  - "Hashtag x Hashtag – co-occurrence" = frequency of two hashtags in same document

Step 4: Create Shared Concept Network

- This is the network for document to documents or authors to authors such that the link is the number of concepts they have in common.
- ORA Twitter importer will create:
  - Author x author – shared hashtags
Step 5: Pruning Big Data

- Handling really big data. You may want to prune what concepts you are using, what documents or authors, and/or what links.

Step 5: Concept Pruning...

- In most data sets most words are used only once. It is generally advisable to remove all words that are not used by some fraction of the authors or in some fraction of texts. For example, only keep words used by at least three authors.
- Many words are highly used but don’t provide a good understanding of the overall themes. These are often referred to as stop words. You probably want to remove all stop words.
- TFIDF is sometimes used to prune concepts.
Step 5: Concept Pruning ... 

- Load the Snowball data which is Twitter data about NATO and Russia in the Baltic region.
- There are 12020 hashtags
- There are 13450 agents
- Select the Agent x Hashtag network
- Click the **Binary link values** checkbox

---

Step 5: Concept Pruning ...

- Use Display Options to show row and column sums
- The column sums give the number of distinct agents that use hashtag
- “Distinct Agents” because we made link weights binary
Step 5: Concept Pruning …

- Sort the columns from high to low by column sum
- Note that NATO, PiS, Trump, Russia, Poland are used the most
- We will reduce the data size by removing hashtags not used by many agents

- We will reduce the data size by removing hashtags not used by many agents
- Use the Column Nodes Select columns by sum (in-degree)
- Dialog appears showing the distribution of agents using hashtags
- Min is 1
- Min + Stddev \( \sim \) 10
- Select these “low” degree hashtags
Step 5: Concept Pruning...

- There are 11469 columns (that is, hashtags) selected.
- These all are used by 10 or fewer agents.
- Use Nodes \ Delete selected nodes to remove the hashtags.
- 551 hashtags remain.

Step 5: Actor or Document Pruning

- If you are mostly interested in the discussion core – you will want to remove all documents or actors that do not share a certain fraction of concepts with other documents or authors. For example, they share less than three concepts.
Step 5: Actor or Document Pruning...

- Create an Agent x Agent – shared hashtag network
- Ensure that Binary link values is checked, since we want to count distinct shared hashtags
- Fold the Agent x Hashtag using the default settings in the dialog that appears

- Resulting Agent x Agent – shared hashtag network is added to the main interface
- Note the link value distribution: at least one pair of nodes has 360 shared hashtags
- We don’t want self-loops, so we will remove them next
Step 5: Actor or Document Pruning...

- Remove self-loops using the Trim \ Remove self-loops menu item

- To find the links with the highest values, do the following:
  - Use Display Options \ Change to list view
  - Use Display Options \ Node Titles
  - Click on the Value column to sort descending by value
Step 5: Actor or Document Pruning…

- Suppose we want to find the super-connected groups within the Agent x Agent – shared hashtags
- These are agents talking about the same things
- Use the menu Trim \ Remove nodes by k-core

Click compute on the Filter Nodes by K-core dialog
- The K-cores are computed
- A K-core contains nodes that all have degree $\geq k$
- K-cores are subsets of each other
- The larger the K, the more shared hashtags
- Our largest core has 290 fully connected nodes! (density = 1)
Step 5: Actor or Document Pruning...

- Click compute on the Filter Nodes by K-core dialog
- The K-cores are computed
- A K-core contains nodes that all have degree >= k
- K-cores are subsets of each other
- The larger the K, the more shared hashtags
- Our largest core has 290 fully connected nodes! (density = 1)

Step 5: Actor or Document Pruning...

- Select a k-core
- The largest k-selected determines what will be kept
- Click Retain K-core to reduce the agents to this core agent group
Step 5: Link pruning

- We could also remove links from the Semantic Network (concept x concept)
- Usually trim out weaker links
- Select the Hashtag x Hashtag network

Within the editor, choose menu Trim \ Remove links by value
- We see that the Avg link value = 3.47 with Std.dev = 14
- We will remove links below the avg + 1 stddev (17)
Step 5: Link pruning...

- Only 680 links remain
- And there are now 11699 isolates
- And 4 larger components

Step 5: Prune by Component

- Click on the network editor
- Use the Trim \ Remove nodes by component to view the components in the network
Step 5: Prune by Component...

- We created many isolates
- 15 dyads
- 5 triads
- We can select which components we want to remove and then click the Remove Nodes button

Step 6: Identify Concepts with High Rhetorical Power

- Having trimmed our data (recall that trimming is optional), we can work with it to find concepts and authors of interest

- Concepts high in one or more dimensions below have high rhetorical power
  - Frequency of use – This is measured as the number of documents/authors that use this concept – these are Hot Topics
  - Connection to other concepts – This is the total degree centrality of the concept in the semantic network
  - Evokability and Invokability – This is the k-betweenness in the semantic network where k = 2
Step 6: Identify Concepts with High Rhetorical Power...

- We will run the Communicative Power report to find the concepts high in rhetorical power.

- Click next and use the given settings.
Step 6: Identify Concepts with High Rhetorical Power...

- The report shows the different categories into which the concepts are categorized
- And lists the top scoring concepts for each category
- It also shows the authors that most frequently use concepts in that score high in
  - Frequency
  - Evokability/invokability
  - Connections to other concepts

Concept Analysis Overview

Concepts are classified according to whether they have high and low values for the measures below. A high value is one in the top third, and a low value is one in the bottom third. Measure values are scaled to be between zero and one.

Connections to other concepts total degree centrality in the semantic network using link values.

Evokability/invokability k-betwenness centrality in the semantic network using inverted link values and computed on nodes no more than 2 links away.

Frequency of Use the number of times the concept is used in the document, which can be weighted.

<table>
<thead>
<tr>
<th>Class</th>
<th>Number of concepts</th>
<th>Percent of concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Symbol</td>
<td>74</td>
<td>0.62%</td>
</tr>
<tr>
<td>Placeholder</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Stereotype</td>
<td>232</td>
<td>1.93%</td>
</tr>
<tr>
<td>Entities</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Allusions</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Buzzword</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Factoid</td>
<td>109</td>
<td>0.91%</td>
</tr>
<tr>
<td>Ordinary Word</td>
<td>2</td>
<td>0.02%</td>
</tr>
</tbody>
</table>
Step 6: Identify Concepts with High Rhetorical Power...

*Author Usage - High Evokability Concepts*

This shows authors that are the most concepts with high evokability.

If the node of author has a higher than normal value (greater than 1 standard deviation) above the normal the row is colored red. The row is green if the node is within 1 standard deviation of the mean. Finally, the row is colored blue if the node has a lower than normal value (less than one standard deviation) below the mean.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Spain</td>
</tr>
<tr>
<td>2</td>
<td>Russia</td>
</tr>
<tr>
<td>3</td>
<td>UK</td>
</tr>
<tr>
<td>4</td>
<td>USA</td>
</tr>
<tr>
<td>5</td>
<td>Britain</td>
</tr>
<tr>
<td>6</td>
<td>NATO</td>
</tr>
<tr>
<td>7</td>
<td>Spain</td>
</tr>
<tr>
<td>8</td>
<td>Pakistan</td>
</tr>
<tr>
<td>9</td>
<td>EU</td>
</tr>
</tbody>
</table>

Step 7: Joint Conversation

- Identify those actors who are influential in the shared concept network. Run the key-entity report and see who has high total degree and betweenness centrality.
Step 7: Joint Conversation...

- Choose the Key-Entity report
- We will need the Agent × Agent – shared hashtag network that we created by folding (see above)

The Key-Entity report will run measures for all nodesets and networks in the meta-network
- Suppose we want only to analyze the Agent × Agent – shared hashtags network
- Use the Filter Data tab on the left-hand side
- Then select to analyze only the agent nodeset and its network
Step 7: Joint Conversation...

- Also, we need to choose only fast measures

Total Degree Centrality

Individuals or organizations who are "in the know" are those who are linked to many others and so by virtue of those connections access to the same thoughts, beliefs of many others. Individuals who are "in the know" are identified by degree centrality in the email and social networks. Those who are ranked high on the list have more connections to others in the same network. The scientific name for this measure is total degree centrality which is calculated on the agent by agent matrix.

If the node of interest has a higher than normal value (greater than 1 standard deviation) above the mean) the icon is colored red. There is a proof of the node is within 1 standard deviation of the mean. Finally, the other a colored blue if the node is less than normal value (less than one standard deviation) below the mean.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Agent</th>
<th>Value</th>
<th>Unusual</th>
<th>Content*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>unimportant</td>
<td>3.098e-05</td>
<td>3.363</td>
<td>-2.958</td>
</tr>
<tr>
<td>2</td>
<td>Washburn_PO</td>
<td>4.917e-05</td>
<td>3.462</td>
<td>-2.972</td>
</tr>
<tr>
<td>3</td>
<td>SpecPhone</td>
<td>9.513e-05</td>
<td>3.363</td>
<td>-2.964</td>
</tr>
<tr>
<td>4</td>
<td>AHHHR</td>
<td>4.692e-05</td>
<td>0.931</td>
<td>0.019</td>
</tr>
<tr>
<td>5</td>
<td>Kelter370</td>
<td>6.912e-05</td>
<td>0.200</td>
<td>-0.048</td>
</tr>
<tr>
<td>6</td>
<td>Jiamaowakufu</td>
<td>9.460e-05</td>
<td>9.294</td>
<td>-1.163</td>
</tr>
<tr>
<td>7</td>
<td>Os_yk_wakufu</td>
<td>6.579e-05</td>
<td>0.834</td>
<td>-1.157</td>
</tr>
<tr>
<td>8</td>
<td>transaction</td>
<td>4.215e-05</td>
<td>0.456</td>
<td>-2.182</td>
</tr>
<tr>
<td>9</td>
<td>jaccab86</td>
<td>4.208e-05</td>
<td>0.346</td>
<td>-2.152</td>
</tr>
<tr>
<td>10</td>
<td>Defsmakufu</td>
<td>3.184e-05</td>
<td>6.746</td>
<td>-2.205</td>
</tr>
</tbody>
</table>

The top ranked nodes are often using the same hashtags.
Step 8: Find “Topics”

- Analyze the Agent x Hashtag network to find groups of hashtags and groups of agents
- This is called topic analysis
- We will run the Topic Analysis report and the algorithms LDA (Latent Dirichlet Analysis) and LSA (Latent Semantic Analysis)

Step 8: Find “Topics”...

- Choose the Topic Analysis report
- We will need the Agent x Agent – shared hashtag network that we created by folding (see above)
Step 8: Find “Topics”…

- Use the Agent x hashtag network as the “Document” x Concept network
- Here we abuse terminology and call Agents Documents because we are following the terminology of topic analysis

This report finds “topics” that are unnamed, but examining the concepts in the topic we can deduce the commonality between them
Step 8: Find “Topics”...

- Also for the topic we get the agents most associated with the topic group

<table>
<thead>
<tr>
<th>Rank</th>
<th>Text</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>mappa plug</td>
<td>0.114</td>
</tr>
<tr>
<td>2</td>
<td>Wealden PR</td>
<td>0.750</td>
</tr>
<tr>
<td>3</td>
<td>NaMa</td>
<td>0.726</td>
</tr>
<tr>
<td>4</td>
<td>spask plug</td>
<td>0.098</td>
</tr>
<tr>
<td>5</td>
<td>Oxted/Wealden</td>
<td>0.659</td>
</tr>
<tr>
<td>6</td>
<td>MB_bolton</td>
<td>0.480</td>
</tr>
<tr>
<td>7</td>
<td>Tyke, neavy</td>
<td>0.270</td>
</tr>
<tr>
<td>8</td>
<td>eg. add</td>
<td>0.240</td>
</tr>
<tr>
<td>9</td>
<td>SW_Lancaster</td>
<td>0.200</td>
</tr>
<tr>
<td>10</td>
<td>borough plug</td>
<td>0.580</td>
</tr>
</tbody>
</table>

Step 9: Compare

- Compare documents or authors by having each of their semantic networks separately and then finding the lossy intersection (that semantic network s.t. each pair of concepts are jointly used by at least 50% of the documents/authors).
- The examine the distribution of the different documents/authors semantic networks from the mean.
- Use the semantic network report to generate this distribution and the lossy intersection. For some corpae you may have to use fewer than 50%
- We can also use the Network Comparison report.
- We need as input the separate meta-networks (one per text) output from Automap
Step 9: Compare...

- A collection of meta-networks output from Automap have been loaded
- Each has a Concept nodeset and a Semantic Network

Step 9: Compare...

- We will select the Semantic Network report
- And run on all meta-networks
Step 9: Compare...

- Use the Concept nodeset
- And not the Central Network values at the bottom

Illustrative related questions

- How can you characterize the discussion in a community?
- How can you assess the potential of a message to reach a wide audience?
- What kinds of messages are most likely to be agreed with?
- What concepts are critical in a discussion?
- What concepts are most powerful when used in a message?