Topic Group and Semantic Network Reports

Jeff Reminga

The CASOS Center
COS Program, School of Computer Science, Carnegie Mellon
Summer Institute 2020

Topic Group Report

- 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)
Example data: Twitter

- 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

Example data: Twitter

- Import the Flood data using the Data Import Wizard
- The data is included in the SI data folder
- The Topic Group report finds and analyzes groups of agents that:
  - Interact amongst themselves (Agent interaction): for example, agents retweeting, replying, quoting, mentioning amongst themselves
  - Shared concepts: tweeting using the same hashtags
• Select the Flood meta-network
• Click the Generate Reports button
• Select the Topic Group report and Click Next

• Topic Group report requires us to specify the nodesets for Agent, Document, and Concepts
• Document is Tweet
• And the networks used to form topic groups
• Select the nodesets as shown

• In particular, we chose to use the "Agent x Agent - All Communication" network as agent interaction
• And use the "Agent x hashtag" network

• Note how the Minimum topic group size we adjusted to 5 (the default is 3)
Topic Group Report...

- Click Next
- ORA will compute the topic groups and display them along with information about each one:
  - Node sizes
  - Network densities
- Select all 5 topic groups
- Click the Extract Topic Group Meta-Networks button to add the topic groups as meta-networks to the main interface

Topic Group Report...

- Click Next and then run the report
- Note that there are 5 new meta-networks added to the main interface
- The agent and hashtag nodesets can be examined to see the topic groups
- In addition, the networks amongst the agents is listed in each network
The report output lists information about the distribution of sizes and densities across the topic groups.

Each topic group is also listed with a hyperlink to a more detailed analysis.
Click a hyperlink for a particular topic group

A table lists the agents that interact the most

Click a hyperlink for a particular topic group

A table lists the agents that interact the most with other agents

Another table lists the agents linked to the most distinct hashtags
Topic Group Report...

Hashing connected to the most distinct Agents (in-degree centrality)

The most distinct agents in the topic group are connected through hashtag centrality. If the node of interest has a higher than normal value (greater than 1 standard deviation above the mean), the node is colored red. If the node is zero, it is within 1 standard deviation of the mean. Finally, the node is colored green if the node has a lower than normal value (less than one standard deviation below the mean).

<table>
<thead>
<tr>
<th>Node ID</th>
<th>HashTag</th>
<th>Value</th>
<th>Uncored</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>FM40198</td>
<td>0.333</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>grouping</td>
<td>0.333</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>Field</td>
<td>0.459</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>Alberta</td>
<td>0.458</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>Australia</td>
<td>0.311</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>ElkhartCity</td>
<td>0.311</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>FM4018</td>
<td>0.311</td>
<td>3</td>
</tr>
</tbody>
</table>

• Most used hashtags are also listed

Semantic Network Report

• The Semantic Network Report will categorize how concepts are used in the agent conversations

• The report requires nodesets agent, document (tweet), and concept (hashtag)

• And networks: Semantic network (Hashtag x Hashtag – Co-occurrence), Document x Concept (Tweet x hashtag), and Agent x Concept (Agent x Concept)
Semantic Network Report

- Click on the Flood dataset
- Click Generate Reports
- And configure the report with the inputs shown
- Click Next twice to run the report

Semantic Network Report...

Communicative Power Analysis

Concepts are classified according to whether they have high and low values for the measures below.

A high value is greater than the 75 percentile value, and a low value is less than the 25 percentile value.

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

Invokability / invokability: betweenness centrality in the semantic network ignoring link values and composed 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>104</td>
<td>8.38%</td>
</tr>
<tr>
<td>Placehold</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Stereotype</td>
<td>41</td>
<td>3.35%</td>
</tr>
<tr>
<td>Entities</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Allusion</td>
<td>4</td>
<td>0.32%</td>
</tr>
<tr>
<td>Buzzword</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Figure</td>
<td>10</td>
<td>0.81%</td>
</tr>
<tr>
<td>Ordinary Word</td>
<td>1</td>
<td>0.08%</td>
</tr>
</tbody>
</table>

- This report analyzes the usage of each hashtag and characterizes its role in the networks
- The hashtag is measured along three dimensions:
  - Frequency
  - Betweenness centrality (invokability)
  - Degree centrality (co-usage with other hashtags)
- A hashtag can be hi/lo in each category
- This yields the 8 categories in the table
Semantic Network Report...

Each category has a table showing the derivation of the scores and the hashtags in the category:

<table>
<thead>
<tr>
<th>Rank</th>
<th>Concept</th>
<th>Connections</th>
<th>Uncoded</th>
<th>Evolvability</th>
<th>Uncoded</th>
<th>Frequency</th>
<th>Uncoded</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>508px</td>
<td>102</td>
<td>0</td>
<td>0</td>
<td>0.073</td>
<td>51</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>WiDestHasTime</td>
<td>4.586e-04</td>
<td>29</td>
<td>0</td>
<td>0.041</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Riots</td>
<td>0.001</td>
<td>81</td>
<td>0</td>
<td>0.038</td>
<td>27</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>GOES16</td>
<td>0.001</td>
<td>69</td>
<td>0</td>
<td>0.033</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>ampel</td>
<td>0.001</td>
<td>75</td>
<td>0</td>
<td>0.031</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Kida</td>
<td>0.002</td>
<td>100</td>
<td>0</td>
<td>0.020</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>qidpol</td>
<td>0.001</td>
<td>75</td>
<td>0</td>
<td>0.020</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Biligioli</td>
<td>2.846e-04</td>
<td>18</td>
<td>0</td>
<td>0.026</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>NewAdm</td>
<td>1.899e-04</td>
<td>12</td>
<td>0</td>
<td>0.021</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Print0WHRSS</td>
<td>2.846e-04</td>
<td>18</td>
<td>0</td>
<td>0.017</td>
<td>12</td>
<td></td>
</tr>
</tbody>
</table>

Further considerations

- The semantic analysis of larger datasets requires that the user trim the data to focus on agents and hashtags of interest
- The semantic network (concept x concept) quickly becomes dense and unwieldy
- Many agents are not of interest
- The remaining slides detail a workflow of manually trimming datasets and shows in detail the construction and manipulation of the data
- The data in the slides is an example, and any agent/document/concept data is applicable
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.

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
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

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
Concept Pruning ...

- 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 ≈ 10
- Select these "low" degree hashtags

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
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.

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 usingthe default settings in the dialog that appears(211,498),(635,887)
Actor or Document Pruning...

- 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

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

• 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
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 $\geq k$
- 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)
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

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
Link pruning...

- 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)

Only 680 links remain
And there are now 11699 isolates
And 4 larger components
Prune by Component

- Click on the network editor
- Use the Trim \ Remove nodes by component to view the components in the network

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
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 $k = 2$

Identify Concepts with High Rhetorical Power...

- We will run the Communicative Power report to find the concepts high in rhetorical power
Identify Concepts with High Rhetorical Power...

• Click next and use the given settings

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
Identify Concepts with High Rhetorical Power...

**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.

Evocability/avoidability: betweenness 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>Emotions</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>

Identify Concepts with High Rhetorical Power...

**Author Usage - High Evocability Concepts**

This shows authors that use the most concepts with high evocability.

<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>USA</td>
</tr>
<tr>
<td>3</td>
<td>Ukraine</td>
</tr>
<tr>
<td>4</td>
<td>UK</td>
</tr>
<tr>
<td>5</td>
<td>USA</td>
</tr>
<tr>
<td>6</td>
<td>Poland</td>
</tr>
<tr>
<td>7</td>
<td>NATO</td>
</tr>
<tr>
<td>8</td>
<td>Waze</td>
</tr>
<tr>
<td>9</td>
<td>Peru</td>
</tr>
<tr>
<td>10</td>
<td>EU</td>
</tr>
</tbody>
</table>

Showing 1 to 10 of 100 entries  Previous   1   2   3   4   5   6   7   8   9   10   Next
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.

Joint Conversation...

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

- The Key-Entity report will run measures for all nodesets and networks in the meta-network.
- Suppose we want only to analyze the Agent x 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.

Joint Conversation...

- Also, we need to choose only fast measures.
Joint Conversation...

Total Degree Centrality

Individuals or organizations who are 'in the know' are those who are linked to many others and so, by virtue of their position, have access to the ideas, thoughts, beliefs of many others. Individuals who are 'in the know' are attracted by degree centrality in the relevant social network. Those who are ranked high on this measure have more connections to others in the same network. The scientific name of this measure is total degree centrality and it is calculated on the agents by aggregating their centrality values.

Total Degree Centrality is measuring the degree to which agents are using hashtags with others

The top ranked nodes are often using the same hashtags

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)
Find “Topics”...

- Choose the Topic Analysis report
- We will need the Agent x Agent – shared hashtag network that we created by folding (see above)

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
Find “Topics”...

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

<table>
<thead>
<tr>
<th>Rank</th>
<th>Concept</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>UE</td>
<td>0.005</td>
</tr>
<tr>
<td>2</td>
<td>Syria</td>
<td>0.005</td>
</tr>
<tr>
<td>3</td>
<td>Bosnia</td>
<td>0.004</td>
</tr>
<tr>
<td>4</td>
<td>Kosovar</td>
<td>0.004</td>
</tr>
<tr>
<td>5</td>
<td>Ukraine</td>
<td>0.004</td>
</tr>
<tr>
<td>6</td>
<td>Bulgaria</td>
<td>0.003</td>
</tr>
<tr>
<td>7</td>
<td>USA</td>
<td>0.003</td>
</tr>
<tr>
<td>8</td>
<td>NATO</td>
<td>0.003</td>
</tr>
<tr>
<td>9</td>
<td>Brazil</td>
<td>0.003</td>
</tr>
</tbody>
</table>

Find “Topics”...

Also for the topic we get the agents most associated with the tropic group.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Concept</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ungrouped[*]</td>
<td>0.914</td>
</tr>
<tr>
<td>2</td>
<td>Wildwood_PK</td>
<td>0.750</td>
</tr>
<tr>
<td>3</td>
<td>NaN</td>
<td>0.630</td>
</tr>
<tr>
<td>4</td>
<td>untappd_app</td>
<td>0.596</td>
</tr>
<tr>
<td>5</td>
<td>OneNoteNotes</td>
<td>0.859</td>
</tr>
<tr>
<td>6</td>
<td>MS2.R35[*]</td>
<td>0.665</td>
</tr>
<tr>
<td>7</td>
<td>color�新</td>
<td>0.670</td>
</tr>
<tr>
<td>8</td>
<td>tip.no[*]</td>
<td>0.820</td>
</tr>
<tr>
<td>9</td>
<td>EB86noscore</td>
<td>0.839</td>
</tr>
<tr>
<td>10</td>
<td>sunglasshd[*]</td>
<td>0.550</td>
</tr>
</tbody>
</table>
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

Compare...

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

- We will select the Semantic Network report
- And run on all meta-networks

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?