Entity-centric Topic Extraction and Exploration: A Network-based Approach

Andreas Spitz and Michael Gertz
March 27, 2018 — ECIR 2018, Grenoble

Heidelberg University, Germany
Database Systems Research Group
<table>
<thead>
<tr>
<th>term</th>
<th>score</th>
</tr>
</thead>
<tbody>
<tr>
<td>skripal</td>
<td>0.83</td>
</tr>
<tr>
<td>nerve</td>
<td>0.77</td>
</tr>
<tr>
<td>agent</td>
<td>0.76</td>
</tr>
<tr>
<td>u.k.</td>
<td>0.61</td>
</tr>
<tr>
<td>russia</td>
<td>0.58</td>
</tr>
<tr>
<td>diplomat</td>
<td>0.45</td>
</tr>
<tr>
<td>intelligence</td>
<td>0.43</td>
</tr>
<tr>
<td>poison</td>
<td>0.33</td>
</tr>
<tr>
<td>daughter</td>
<td>0.19</td>
</tr>
<tr>
<td>yulia</td>
<td>0.17</td>
</tr>
</tbody>
</table>
Disadvantages of Traditional (LDA) Topics

Substantial runtime requirements that increase

▶ with the number of documents
▶ with the number of topics
Disadvantages of Traditional (LDA) Topics

Substantial runtime requirements that increase

- with the number of documents
- with the number of topics

Limited flexibility when

- changing the number of topics
- updating the underlying data / processing data streams
Disadvantages of Traditional (LDA) Topics

Substantial runtime requirements that increase

- with the number of documents
- with the number of topics

Limited flexibility when

- changing the number of topics
- updating the underlying data / processing data streams

Limited support for explorations of

- topic labels / topic descriptions
- relations between topics
## Entity-centric Network Topics

<table>
<thead>
<tr>
<th>term</th>
<th>score</th>
</tr>
</thead>
<tbody>
<tr>
<td>skripal</td>
<td>0.83</td>
</tr>
<tr>
<td>nerve</td>
<td>0.77</td>
</tr>
<tr>
<td>agent</td>
<td>0.76</td>
</tr>
<tr>
<td>u.k.</td>
<td>0.61</td>
</tr>
<tr>
<td>russia</td>
<td>0.58</td>
</tr>
<tr>
<td>diplomat</td>
<td>0.45</td>
</tr>
<tr>
<td>intelligence</td>
<td>0.43</td>
</tr>
<tr>
<td>poison</td>
<td>0.33</td>
</tr>
<tr>
<td>daughter</td>
<td>0.19</td>
</tr>
<tr>
<td>yulia</td>
<td>0.17</td>
</tr>
</tbody>
</table>
## Entity-centric Network Topics

<table>
<thead>
<tr>
<th>term</th>
<th>score</th>
</tr>
</thead>
<tbody>
<tr>
<td>skripal</td>
<td>0.83</td>
</tr>
<tr>
<td>nerve</td>
<td>0.77</td>
</tr>
<tr>
<td>agent</td>
<td>0.76</td>
</tr>
<tr>
<td>u.k.</td>
<td>0.61</td>
</tr>
<tr>
<td>russia</td>
<td>0.58</td>
</tr>
<tr>
<td>diplomat</td>
<td>0.45</td>
</tr>
<tr>
<td>intelligence</td>
<td>0.43</td>
</tr>
<tr>
<td>poison</td>
<td>0.33</td>
</tr>
<tr>
<td>daughter</td>
<td>0.19</td>
</tr>
<tr>
<td>yulia</td>
<td>0.17</td>
</tr>
</tbody>
</table>

![Network Diagram](image)
Implicit Entity Networks
What Are Implicit Entity Networks?

What Are Implicit Entity Networks?

What Are Implicit Entity Networks?

Extracting Implicit Networks From Text

**Data Structure**

- **D(e):** Documents in which edge e occurs
- **T(e):** Publication timestamps of documents D(e)
- **Δ(e):** Sentence distances between the nodes of e
- **c(e):** Total number of occurrences of edge e

**Diagram:**
- Annotated document collection
- Implicit network representation

**Nodes:** t₁, e₁, e₂

**Edges:** e₁ → e → e₂
Network Topic Construction
Parallel Edge Aggregation And Ranking

\[ \omega(e) = 3 \cdot \left[ \frac{|D(v_1) \cup D(v_2)|}{|D(e)|} + \frac{\max\{T(e)\} - \min\{T(e)\}}{|T(e)|} + \frac{c(e)}{\sum_{\delta \in \Delta(e)} \exp(-\delta)} \right]^{-1} \]

coverage \hspace{1cm} \text{temporal coverage} \hspace{1cm} \text{distance}

\(D(e)\): documents in which edge \(e\) occurs
\(T(e)\): publication timestamps of documents \(D(e)\)
\(\Delta(e)\): sentence distances between the nodes \(v_1\) and \(v_2\)
\(c(e)\): total number of occurrences of edge \(e\)
Intuition:

- edges between entities correspond to seeds of topics
Intuition:

- edges between entities correspond to seeds of topics
- topics can be grown around seeds by adding relevant terms
Topic Extraction and Triangular Growth

Intuition:

- edges between entities correspond to seeds of topics
- topics can be grown around seeds by adding relevant terms
For a demonstration of entity ranking in implicit networks see:
Topic Overlap and Merging Topics
Topic Overlap and Merging Topics
Topic Overlap and Merging Topics
Overview: News Article Data

English news articles from RSS feeds:

- 14 news outlets (from US, UK, and AU)
- 6 months (Jun 1 - Nov 30, 2016)
- 127,500 articles
- 5,400,000 sentences

The resulting implicit network has:

- 119,300 entities
- 329,000 terms
- 10,600,000 edges
Overview: News Article Data

English news articles from RSS feeds:
- 14 news outlets (from US, UK, and AU)
- 6 months (Jun 1 - Nov 30, 2016)
- 127.5 thousand articles
- 5.4 million sentences

NLP processing pipeline:
- Part-of-speech and sentence tagging: Stanford POS tagger
- Entity classification: YAGO classes (LOC, ORG, PER)
- Named entity recognition and linking:
Overview: News Article Data

English news articles from RSS feeds:
- 14 news outlets (from US, UK, and AU)
- 6 months (Jun 1 - Nov 30, 2016)
- 127.5 thousand articles
- 5.4 million sentences

The resulting implicit network has
- 119.3 thousand entities
- 329.0 thousand terms
- 10.6 million edges

NLP processing pipeline:
- Part-of-speech and sentence tagging: Stanford POS tagger
- Entity classification: YAGO classes (LOC, ORG, PER)
- Named entity recognition and linking:
Network news topics from CNN
June - July 2016
Network Topic Evolution

Network news topics from CNN (2016)

June - July

August - September
Network news topics from June - July 2016

CNN

Guardian

Topics Across Different News Outlets
Comparison to Classic Topics
Term Ranking in Network Topics

\[
\begin{align*}
\text{Term score} & \quad t_1 \\
\omega(e_1, t_1) & \quad \omega(e_2, t_1) \\
\vdots & \quad t_n \\
e_1 & \quad e_2
\end{align*}
\]
Term Ranking in Network Topics

\[
\begin{align*}
\omega(e_1, t_1) & \quad \omega(e_2, t_1) \\
\vdots & \quad \vdots \\
\omega(e_1, t_n) & \quad \omega(e_2, t_n)
\end{align*}
\]

<table>
<thead>
<tr>
<th>term</th>
<th>score</th>
</tr>
</thead>
<tbody>
<tr>
<td>(t_1)</td>
<td>(\min{\omega(e_1, t_1), \omega(e_2, t_1)})</td>
</tr>
<tr>
<td>(t_2)</td>
<td>(\min{\omega(e_1, t_2), \omega(e_2, t_2)})</td>
</tr>
<tr>
<td>(\vdots)</td>
<td>(\vdots)</td>
</tr>
<tr>
<td>(t_n)</td>
<td>(\min{\omega(e_1, t_n), \omega(e_2, t_n)})</td>
</tr>
<tr>
<td>Term</td>
<td>Score</td>
</tr>
<tr>
<td>------------</td>
<td>-------</td>
</tr>
<tr>
<td>syrian</td>
<td>0.14</td>
</tr>
<tr>
<td>rebel-held</td>
<td>0.12</td>
</tr>
<tr>
<td>rebel</td>
<td>0.06</td>
</tr>
<tr>
<td>cease-fir</td>
<td>0.05</td>
</tr>
<tr>
<td>bombard</td>
<td>0.05</td>
</tr>
<tr>
<td>bomb</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Network news topics from the New York Times (Jun - Nov 2016)
Topic Overlap Comparison

- **topic size 5**
- **topic size 10**
- **topic size 50**

### LDA

- network

### Average topic overlap vs number of topics

- **BBC**
- **CBS**
- **CNN**
- **Guardian**
- **IBTimes**
- **Independent**
- **LATimes**
- **NYTimes**
- **Reuters**
- **Skynews**
- **SMH**
- **Telegraph**
- **USAtoday**
- **WPost**
Discussion & Summary
Benefits vs. traditional topics:

- faster extraction than LDA topics
- runtime contained in data preparation
- number of topics is flexible
Benefits of Entity-centric Network Topics

Benefits vs. traditional topics:
- faster extraction than LDA topics
- runtime contained in data preparation
- number of topics is flexible

Stream compatibility:
- document updates require only (sub-) graph updates
Flexibility of Entity-centric Network Topics

Intuitive exploration of topics:

- network visualizations instead of term lists
- entities act as labels for topics
Flexibility of Entity-centric Network Topics

Intuitive exploration of topics:
- network visualizations instead of term lists
- entities act as labels for topics

Efficient support of interactive explorations:
- Adding more topic seeds (edges):
  $O(\log n)$ for edge lookup with index support
- Adding more descriptive terms:
  $O(\langle k \rangle)$ for average node degree $\langle k \rangle$
Summary

Data and implementation are available online:
- [data] Implicit news network
- [code] Implicit network extraction
- [code] Topic exploration and extraction

https://dbs.ifi.uni-heidelberg.de/resources/nwtopics/
Summary

Data and implementation are available online:

- [data] Implicit news network
- [code] Implicit network extraction
- [code] Topic exploration and extraction

https://dbs.ifi.uni-heidelberg.de/resources/nwtopics/