UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

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

A Topic From Recent News

term	score
skripal	0.83
nerve	0.77
agent	0.76
u.k.	0.61
russia	0.58
diplomat	0.45
intelligence	0.43
poison	0.33
daughter	0.19
yulia	0.17

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

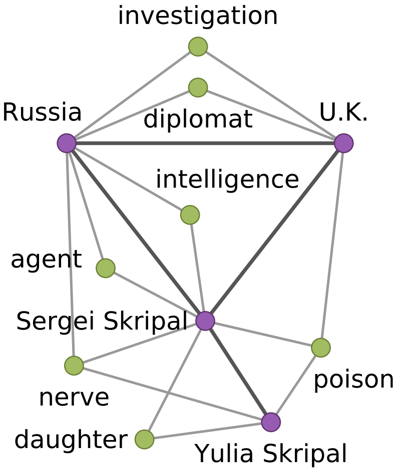
- topic labels / topic descriptions
- relations between topics

Entity-centric Network Topics

term	score
skripal	0.83
nerve	0.77
agent	0.76
u.k.	0.61
russia	0.58
diplomat	0.45
intelligence	0.43
poison	0.33
daughter	0.19
yulia	0.17

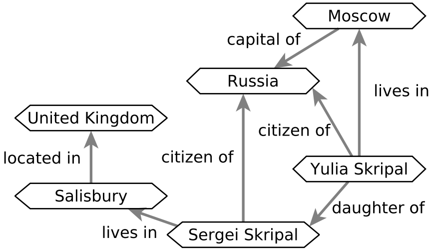
Entity-centric Network Topics

term	score
skripal	0.83
nerve	0.77
agent	0.76
u.k.	0.61
russia	0.58
diplomat	0.45
intelligence	0.43
poison	0.33
daughter	0.19
yulia	0.17



Implicit Entity Networks

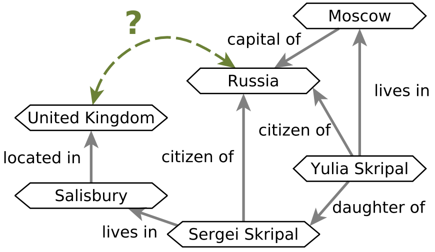
What Are Implicit Entity Networks?



Knowledge Graph

A. Spitz and M. Gertz. "Terms over LOAD: Leveraging Named Entities for Cross-Document Extraction and Summarization of Events". In: *ACM SIGIR*. 2016

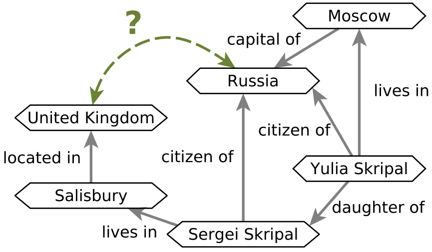
What Are Implicit Entity Networks?



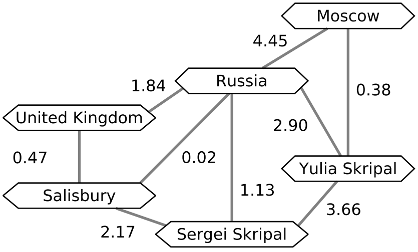
Knowledge Graph

A. Spitz and M. Gertz. "Terms over LOAD: Leveraging Named Entities for Cross-Document Extraction and Summarization of Events". In: *ACM SIGIR*. 2016

What Are Implicit Entity Networks?



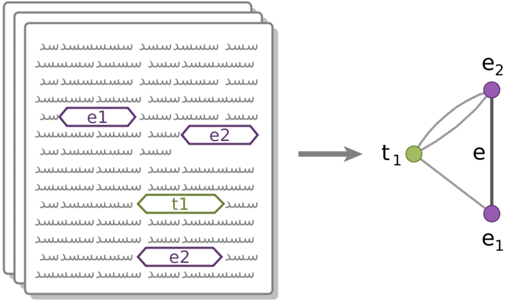
Knowledge Graph



Implicit Network

A. Spitz and M. Gertz. "Terms over LOAD: Leveraging Named Entities for Cross-Document Extraction and Summarization of Events". In: *ACM SIGIR*. 2016

Extracting Implicit Networks From Text



annotated document collection

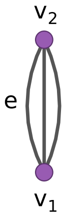
implicit network representation

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

Network Topic Construction

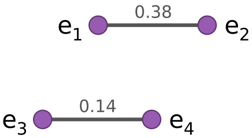
Parallel Edge Aggregation And Ranking

$$\omega(e) = 3 \cdot \Bigg[\underbrace{\frac{|D(v_1) \cup D(v_2)|}{|D(e)|}}_{\text{coverage}} + \underbrace{\frac{\max\{T(e)\} - \min\{T(e)\}}{|T(e)|}}_{\text{temporal coverage}} + \underbrace{\frac{c(e)}{\sum_{\delta \in \Delta(e)} \exp(-\delta)}}_{\text{distance}} \Bigg]^{-1}$$



D(e): documents in which edge e occurs
T(e): publication timestamps of documents D(e)
Δ(e): sentence distances between the nodes v1 and v2
c(e): total number of occurrences of edge e

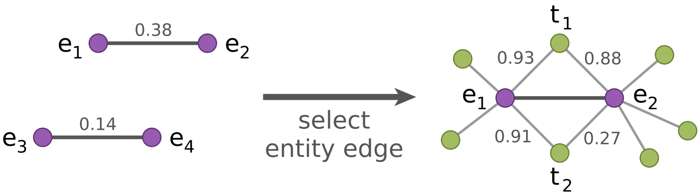
Topic Extraction and Triangular Growth



Intuition:

- edges between entities correspond to seeds of topics

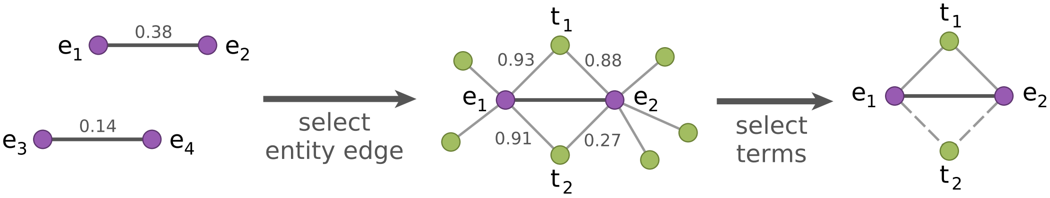
Topic Extraction and Triangular Growth



Intuition:

- edges between entities correspond to seeds of topics
- topics can be grown around seeds by adding relevant terms

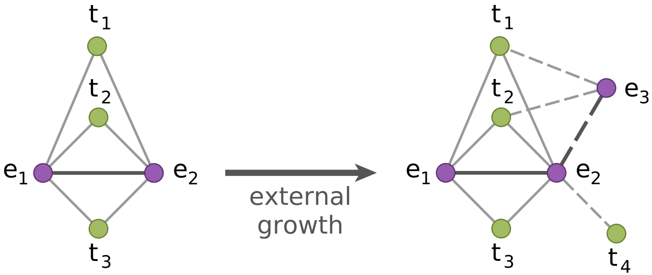
Topic Extraction and Triangular Growth



Intuition:

- edges between entities correspond to seeds of topics
- topics can be grown around seeds by adding relevant terms

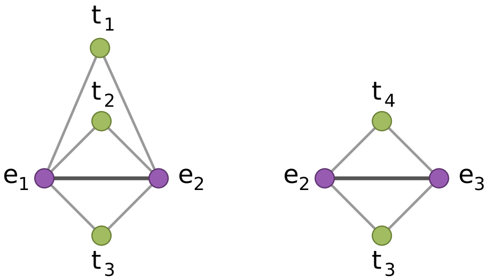
Topic Growth by External Nodes



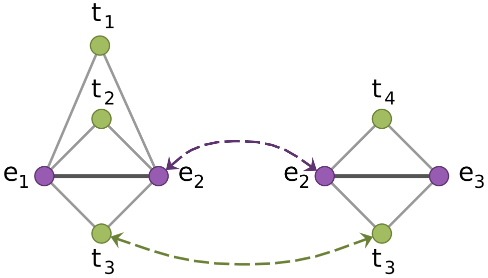
For a demonstration of entity ranking in implicit networks see:

A. Spitz, S. Almasian, and M. Gertz. "EVELIN: Exploration of Event and Entity Links in Implicit Networks". In: *WWW Companion*. 2017. URL: `http://evelin.ifi.uni-heidelberg.de`

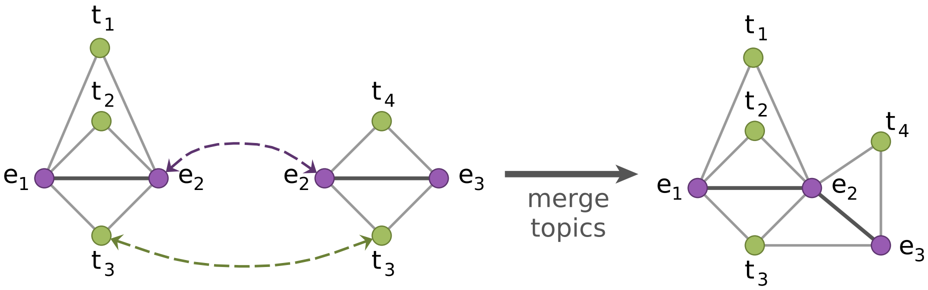
Topic Overlap and Merging Topics



Topic Overlap and Merging Topics



Topic Overlap and Merging Topics



Topic Exploration

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

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:

AMBIVERSE
Text to Knowledge

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

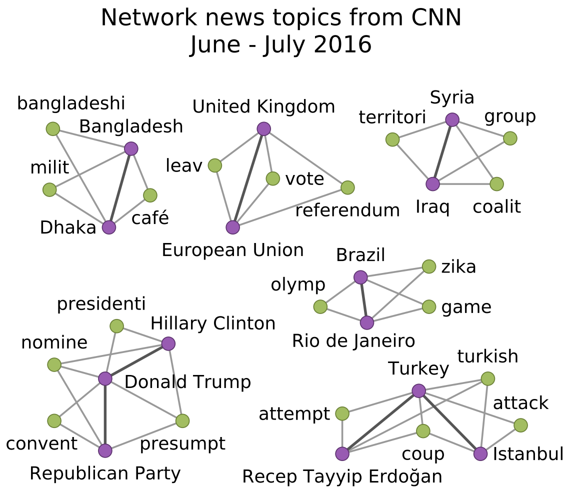
The resulting implicit network has

- 119.3 thousand entities
- 329.0 thousand terms
- 10.6 million edges

NLP processing pipeline:

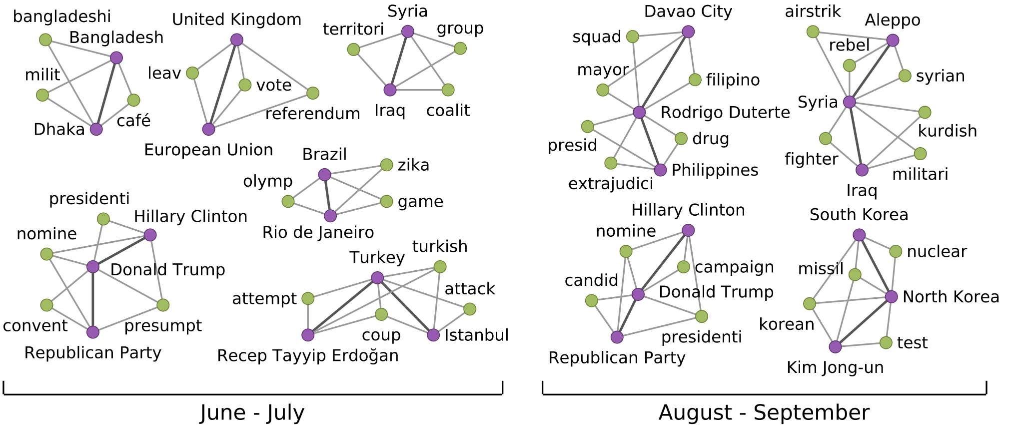
- Part-of-speech and sentence tagging:
  Stanford POS tagger
- Entity classification:
  YAGO classes (LOC, ORG, PER)
- Named entity recognition and linking:

Network Topic Example



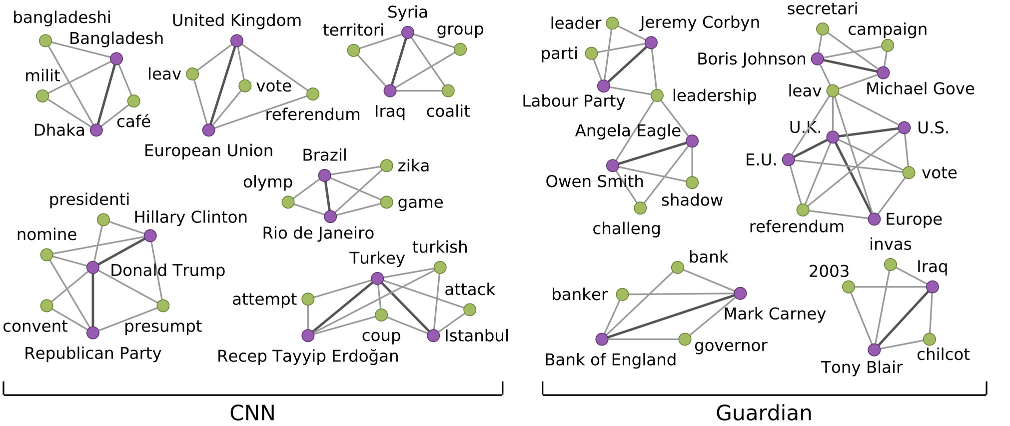
Network Topic Evolution

Network news topics from CNN (2016)



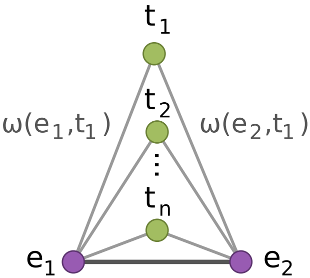
Topics Across Different News Outlets

Network news topics from June - July 2016

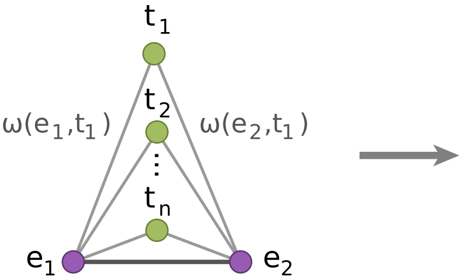


Comparison to Classic Topics

Term Ranking in Network Topics



Term Ranking in Network Topics



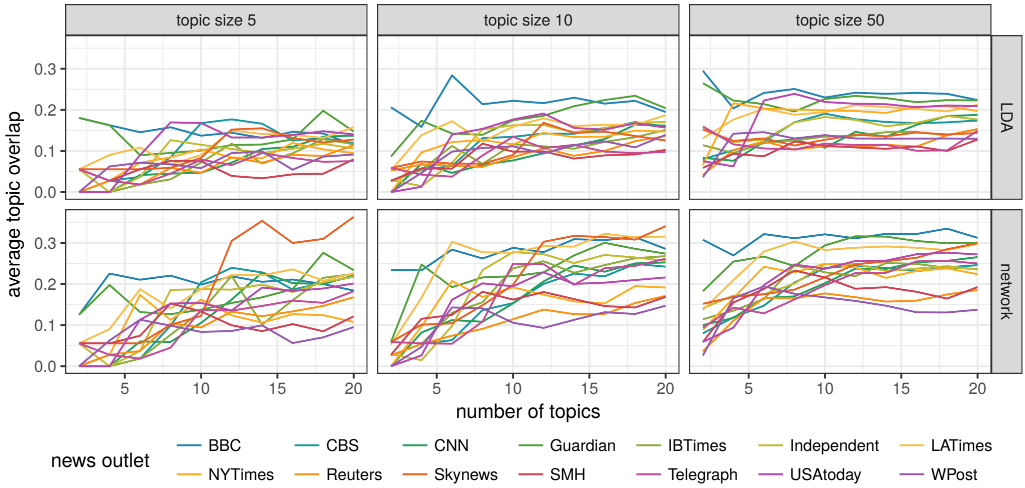
term	score
t_1	$\min\{\omega(e_1, t_1),\ \omega(e_2, t_1)\}$
t_2	$\min\{\omega(e_1, t_2),\ \omega(e_2, t_2)\}$
$\vdots$	$\vdots$
t_n	$\min\{\omega(e_1, t_n),\ \omega(e_2, t_n)\}$

Classic Topics From Network Topics

Beirut - Lebanon Q3820 - Q822		Russia - Moscow Q159 - Q649		Russia - Putin Q159 - Q7747		Trump - Obama Q22686 - Q76	
term	score	term	score	term	score	term	score
syrian	0.14	russian	0.28	russian	0.29	presid	0.40
rebel-held	0.12	soviet	0.06	presid	0.18	american	0.21
rebel	0.06	nato	0.06	annex	0.09	republican	0.19
cease-fir	0.05	diplomat	0.06	nato	0.08	democrat	0.19
bombard	0.05	syrian	0.06	hack	0.08	campaign	0.18
bomb	0.04	rebel	0.05	west	0.08	administr	0.17

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

Topic Overlap Comparison

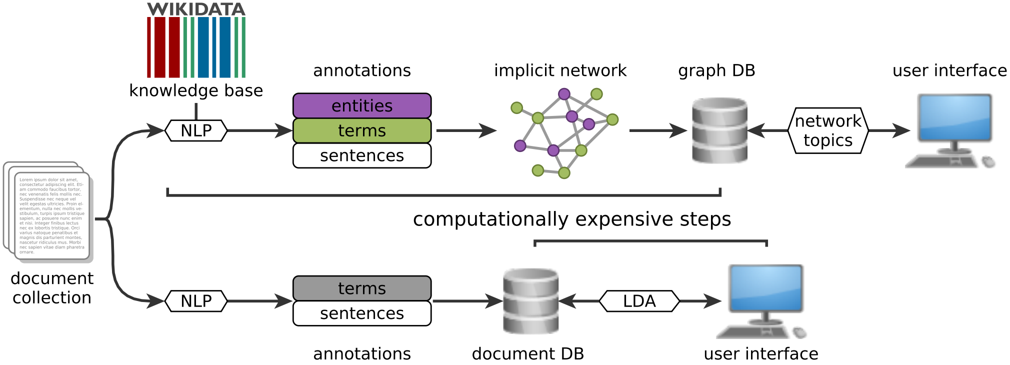


Discussion & Summary

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



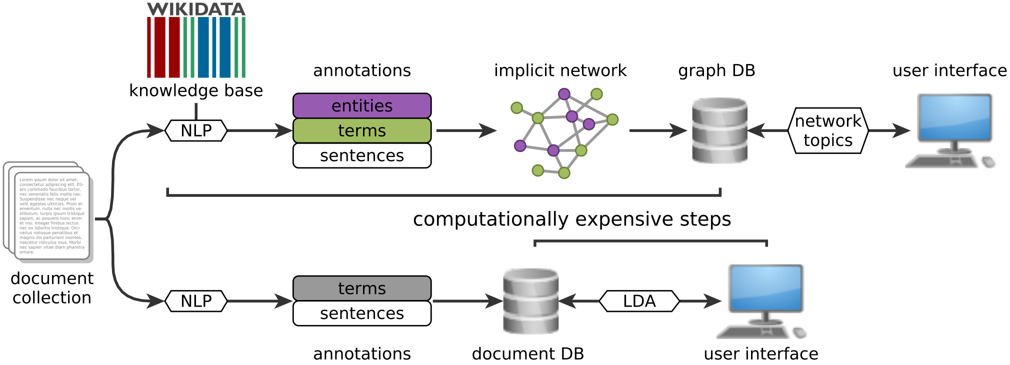
Benefits of Entity-centric Network Topics

Benefits vs. traditional topics:

- faster extraction than LDA topics
- runtime contained in data preparation
- number of topics is flexible

Stream compatibility:

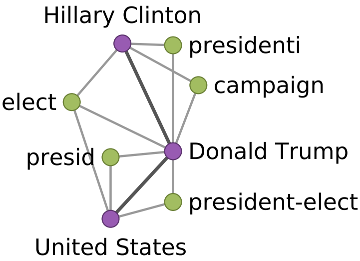
- document updates require only (sub-) graph updates



Flexibility of Entity-centric Network Topics

Intuitive exploration of topics:

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



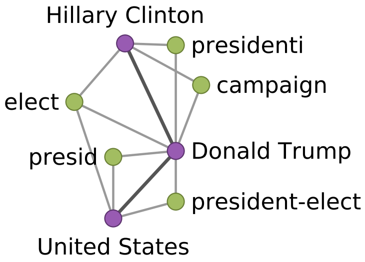
Flexibility of Entity-centric Network Topics

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):
  $\mathcal{O}(\log n)$ for edge lookup with index support
- Adding more descriptive terms:
  $\mathcal{O}(\langle k \rangle)$ 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/