Exploring Entity-centric Networks in Entangled News Streams

Andreas Spitz and Michael Gertz
April 25, 2018 — WWW 2018, Lyon

Heidelberg University, Germany
Database Systems Research Group
Parallel News Streams

- the guardian
- The New York Times
- The Washington Post

(time)}
DON'T CROSS THE STREAMS

IT WOULD BE BAD
Entangled News Streams

Core idea: entity cooccurrences characterize stitching points between news streams

- the guardian
- The New York Times
- The Washington Post

(time) → context(s)
Entangled News Streams

Core idea: entity cooccurrences characterize stitching points between news streams
Entangled News Streams

Core idea: entity cooccurrences characterize stitching points between news streams
Implicit Entity Networks
Implicit Network Aggregation

Following an investigation, the FBI recommends no charges be filed against Hillary Clinton over her emails.

Implicit Networks of Text Streams
The FBI investigates the use of a private email server by Hillary Clinton, Secretary of State.
The FBI investigates the use of a private email server by Hillary Clinton, Secretary of State.
Context-based Aggregation of Edges

The FBI investigates the use of a private email server by Hillary Clinton, Secretary of State.

Time stamp $\tau = 2015-08-12$

Following an investigation, the FBI recommends no charges be filed against Hillary Clinton over her emails.

Time stamp $\tau = 2016-07-05$

$T = \{2015-08-12, 2016-07-05\}$

$\omega = 2$ (importance)

$\lambda = 2$ (multiplicity)
Edge Aggregation Approaches

Streaming aggregation:
- Compare similarity of new edge \((v, w, \cdot)\) to existing edges \((v, w, \cdot)\).
- If similarity threshold is exceeded:
  - Merge with existing edge.
- Otherwise, insert as new parallel edge.

Static aggregation / clustering:
- Collect all parallel edges.
- Cluster parallel edges (density-based).
- Discard “noisy” edges.
- Aggregate edges within clusters.
Edge Aggregation Approaches

Streaming aggregation:
- Compare similarity of new edge 
  \((v, w, \cdot)\) to existing edges \((v, w, \cdot)\)
- If similarity threshold is exceeded:
  merge with existing edge
- Otherwise, insert as new parallel edge

Static aggregation / clustering:
- Collect all parallel edges
- Cluster parallel edges (density-based)
- Discard “noisy” edges
- Aggregate edges within clusters
Edge Aggregation Approaches

Streaming aggregation:
- Compare similarity of new edge \((v, w, \cdot)\) to existing edges \((v, w, \cdot)\)
- If similarity threshold is exceeded: merge with existing edge
- Otherwise, insert as new parallel edge

Static aggregation / clustering:
- Collect all parallel edges
- Cluster parallel edges (density-based)
- Discard “noisy” edges
- aggregate edges within clusters
Application Examples
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
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
- Temporal tagging: HeidelTime
- Entity classification: YAGO classes (LOC, ORG, PER)
- Named entity recognition and linking:
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
- 125 thousand entities
- 351 thousand terms
- 83.4 million edges

NLP processing pipeline:
- Part-of-speech and sentence tagging: Stanford POS tagger
- Temporal tagging: HeidelTime
- Entity classification: YAGO classes (LOC, ORG, PER)
- Named entity recognition and linking:
Evolution of Entity Contexts

Topics for David Cameron (Q192) – UK (Q145)

- Brexit nation favour
- Demand govern
- Referendum Ukip vote
- Westminster campaign
- Prime minister leader resign pro-Brexit

Relative frequency of mentions

June to October
Andreas Spitz and Michael Gertz. “Entity-Centric Topic Extraction and Exploration: A Network-Based Approach”. In: ECIR. 2018
Further Applications

News analysis and exploration:

- Contrastive source comparison
- Coverage bias
- Evolution of news stories
- Event description
- ...

NLP and IR applications:

- Entity disambiguation
- (Extractive) summarization
- Relationship extraction
- ...
Further Applications

News analysis and exploration:
- Contrastive source comparison
- Coverage bias
- Evolution of news stories
- Event description
- ...

NLP and IR applications:
- Entity disambiguation
- (Extractive) summarization
- Relationship extraction
- ...

13
Resources
Resources

Data and implementation are available online:

- [data] Implicit news stream network
- [code] Implicit network extraction
- [code] Entity query and topic extraction

https://dbs.ifi.uni-heidelberg.de/resources/newsstream/
Resources

Data and implementation are available online:

- [data] Implicit news stream network
- [code] Implicit network extraction
- [code] Entity query and topic extraction

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