TopExNet:
Entity-centric Network Topic Exploration in News Streams

Andreas Spitz, Satya Almasian and Michael Gertz

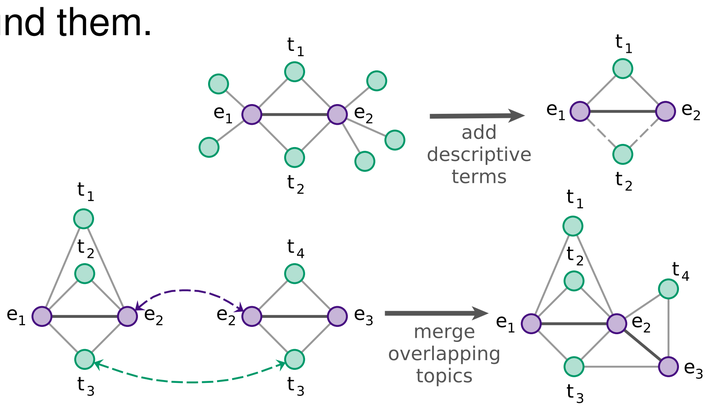
Database Systems Research Group, Heidelberg University, Im Neuenheimer Feld 205, 69120 Heidelberg, Germany

The Idea: Network Topics

Traditional topic models generate ranked lists of words that can be difficult to interpret. Such topics are often static and costly to extract, which impedes the exploration of topics in dynamic collections or streams of documents. We demonstrate **entity-centric network topics** as an alternative solution to extracting topics, which enables an **interactive and visual exploration** of topics contained in **implicit network** representations of documents.

Network Topic Extraction

Topics tend to focus on entities and their relations. In news, these are persons, locations, or organizations. Thus, we consider implicit relations between entities as **seeds of topics** and include terms to grow descriptive subgraphs around them.



Topics around individual seed relations are merged to create larger topics if they overlap.

Database Structure

For the selection of date ranges and specific news outlets by the user during the interactive extraction of network topics, an efficient graph representation is necessary. We use:

- Separate edge lists of entity-entity and term-entity cooccurrences
- Partial edge aggregation of edges by news outlet and date
- Lookup tables for node occurrence statistics
- A relational database architecture and index

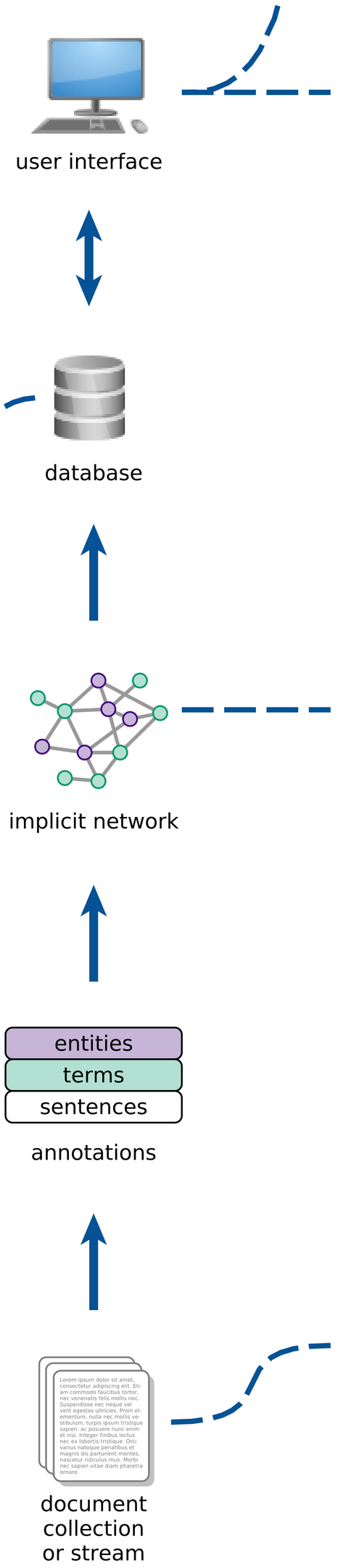
Document Processing and Annotation

Documents are processed individually as they occur in the stream, including:

- Sentence detection and tokenization
- Named entity annotation (persons, locations, organizations)
- Entity linking to Wikidata
- Annotation of non-entity tokens as terms
- Metadata extraction of the publication date and the news outlet

Try it Yourself!

The interface for the 2016 news article data set is available as a Website that you can use from most mobile devices: `https://topexnet.ifi.uni-heidelberg.de`



How-To: Query Input

The interface supports two primary modes of searching for initial topics:

Entity-specific topic search:
Select a date range, a set of news outlets, and exactly two query entities, then search. The topic for this pair of entities is constructed.

Global topic search:
Select a date range, a set of news outlets, and no query entities, then search. The topics for the top-ranked seed edges are constructed.

How-To: Interactive Topic Exploration

Network topics can also be extended by entity-centric ranking of adjacent nodes [2].

Adding related entities:
Select a single entity and right click (touch and hold) to select related entities or delete an entity. Add terms to edges by right clicking (touching and holding) on the edge or on the canvas.

Knowledge base links:
Select a single entity and right click (touch and hold), then follow the menu to Wikidata.

Links to source articles:
Select multiple entities and right click (touch and hold) for a list of related news articles.

Implicit Network Extraction

Conceptual approach:
Extract a network of entities and terms that represents implicit entity relations and their context in the documents [1].

Edge weighting:
We generate weights ω for edges $e = (v, w)$ as

$$\omega(e, \tau) = 3 \left[\frac{|D_v \cup D_w|}{|D_e|} + \frac{\tau_1 - \tau_2}{|\mathcal{T}_e|} + \frac{D_{max}}{\Delta_e} \right]^{-1}$$

where D_v, D_w, and D_e denote the sets of documents in which v, w, and e occur, $\tau = (\tau_1, \tau_2)$ is a date interval, and $\Delta_e = \sum_d \exp(-\delta(v, w, d))$ is the sum of decaying reciprocal distances δ between mentions of v and w in documents d.

News Article Data Set

We collect news articles from the RSS feeds of 14 international English-speaking news outlets, and remove the boilerplate.

- Focus on political news
- Outlets from the U.S., U.K., and Australia
- Time frame: June 1 to November 30, 2016

References

[1] Andreas Spitz and Michael Gertz. **Entity-centric Topic Extraction and Exploration: A Network-based Approach**. 2018, *ECIR'18*

[2] Andreas Spitz, Satya Almasian and Michael Gertz. **EVELIN: Exploration of Event and Entity Links in Implicit Networks**. 2017, *WWW'17*

Contact Information:

Andreas Spitz
spitz@informatik.uni-heidelberg.de
`http://dbs.ifi.uni-heidelberg.de/`

This demonstration was presented at the 12th International Conference on Web Search and Data Mining (WSDM '19), February 11-15, 2019, Melbourne, Australia.