Entity-Centric Topic Extraction and Exploration: A Network-Based Approach

Andreas Spitz(✉) and Michael Gertz

Heidelberg University, Heidelberg, Germany
{spitz,gertz}@informatik.uni-heidelberg.de

Abstract. Topic modeling is an important tool in the analysis of corpora and the classification and clustering of documents. Various extensions of the underlying graphical models have been proposed to address hierarchical or dynamical topics. However, despite their popularity, topic models face problems in the exploration and correlation of the (often unknown number of) topics extracted from a document collection, and rely on compute-intensive graphical models. In this paper, we present a novel framework for exploring evolving corpora of news articles in terms of topics covered over time. Our approach is based on implicit networks representing the cooccurrences of entities and terms in the documents as weighted edges. Edges with high weight between entities are indicative of topics, allowing the context of a topic to be explored incrementally by growing network sub-structures. Since the exploration of topics corresponds to local operations in the network, it is efficient and interactive. Adding new news articles to the collection simply updates the network, thus avoiding expensive recomputations of term and topic distributions.

Keywords: Networks · Topic models · Evolving networks

1 Introduction

Given a collection of time-stamped documents spanning a period of time, what topics are covered in the documents and how do they evolve? This question is common in corpus analysis, ranging from the exploration of news media to the analysis of corpora of historic documents. In most cases, answers to the above questions are provided by employing probabilistic topic models, which are based on Latent Dirichlet Allocation, LDA [4]. In these models, documents are assumed to consist of one or more topics, which are distributions of words. Once identified, the topics for a corpus can also be used for document classification and clustering. Due to this range of applications, a multitude of tools for computing topics and extensions to LDA have been proposed (for an overview, see [2]).

Despite the versatility of topic models, topics are often simply represented as lists of ranked terms, some of which are even difficult to associate with the topic. More recently, there have thus been approaches that enable the exploration and analysis of the computed topics and ranked terms (e.g., [7,11]). To utilize

G. Pasi et al. (Eds.): ECIR 2018, LNCS 10772, pp. 3–15, 2018.
https://doi.org/10.1007/978-3-319-76941-7_1

additional annotation data during topic extraction, some approaches also include named entities in the topic model [9,12], an aspect that is important for the analysis of news articles, which typically revolve around such entities. However, these approaches rely on computationally expensive graphical models, and often produce a fixed number of topics that cannot be altered after the extraction.

In this paper, we present an alternative and efficient framework for the exploration and analysis of topics, including their evolution over time. Our approach is based on *implicit networks* that are constructed from the cooccurrences of terms and named entities in the documents, which are encoded as weighted edges in the network. Based on the central role that entities play in the evolution of news stories, we conjecture that frequently cooccurring pairs of entities are indicative of a topic. Starting from such seed edges between two words in the network representation, one can then construct the *context* of a potential topic by following other highly weighted edges to adjacent terms and entities. With this approach, the exploratory character of topic discovery and the overlap between identified topics becomes obvious. Important seed edges can easily be determined or expanded, meaning that the model is not constrained to a fixed number of topics. For news articles, publication dates provide an effective means of focussing on entity and term cooccurrences in a given time frame. Thus, the evolution of topics in terms of edge weights and word contexts can be effectively explored in such an implicit network, which also supports the addition of new documents, resulting in new nodes, edges, and updated edge weights. Most importantly, it is not necessary to recompute topics and word distributions for an evolving corpus when new documents are added. As a result, a network-based framework can support a variety of entity-centric topic analysis and exploration tasks.

2 Related Work

The framework that we propose in this paper is related to two broader areas of research, namely topic modeling and network-based document representation.

Topic Modeling. Since the introduction of topic models based on Latent Dirichlet Allocation [4], numerous frameworks for topic models have been proposed. The review by Blei summarizes the diverse approaches and directions [2]. However, since topic models primarily extract topics from a document collection in the form of ranked lists of terms, it has been questioned to what extent such a representation is semantically meaningful or interpretable, beyond providing an approximate initial parameter of topics to discover. This issue has been addressed by Chang et al. [5] by proposing novel quantitative methods for evaluating the semantic meaning of topics. Our work is most closely related to entity-centric topic modeling, e.g., [9,12], which focusses on (named) entities and their inclusion in the topics. These approaches rely on the original concept of LDA and thus share the inherent problems of divining an appropriate number of topics to discover and the necessity of interpreting lists of ranked terms.

More recent works in the area of topic modeling also propose frameworks that enable the interactive construction and exploration of topics from a collection of documents, e.g., [7,11], thus providing the user with added flexibility in terms of corpus analysis. Another interesting direction that implicitly addresses the aspect of collocation is the combination of word embeddings and topic models [15]. Common to most extensions of probabilistic models for entity-centric topic modeling and exploration is the reliance on a prior (computationally expensive) construction of lists of ranked terms. In contrast, we use an underlying network document model that adds versatility to the subsequent steps.

Network-based Document Models and Analyses. Network-based representations of documents and corpora have become a popular tool for analyzing the context of entities and terms. The most prominent and well-known examples are collocation networks (for an in-depth discussion, see [6,8]), even though the derived networks are only used for representation and not for network analysis tasks. Some recent works employ network-based representations of documents for the discovery and description of events [14,16,18]. These works focus on named entities in the documents, for which an implicit network is constructed from entity and term cooccurrences, but do not consider the extraction, exploration, or temporal evolution of topics. Other recent approaches in support of more general information retrieval tasks such as ranking have been proposed by Blanco and Lioma [1] and Rousseau and Vazirgiannis [13].

A combination of term collocation patterns and topic models was recently proposed by Zuo et al. [20]. Compared to the network-based approach that we present here, their approach is tailored towards short documents and relies on LDA, thus incurring the same problems as the topic models outlined above.

3 Network Model

In the following, we describe how an implicit network representation is constructed for a collection $D = \{d_1, \ldots, d_n\}$ of news articles. We assume each document d_i to have a timestamp $t(d_i)$, indicating the article's publication time. The network for the collection D, denoted as a graph $G_D(V, E)$, consists of a set of nodes V representing words in the documents and a set of labeled edges E that describe the *cooccurrence* of pairs of words in the documents.

Network Nodes. With the focus on entity-centric topics, we distinguish different types of entities such as persons, organizations, and locations. We assume that words in the sentences have been tagged with respect to a known set of named entities NE. Stop words are removed, the remaining untagged words are denoted as the set of terms T. We then let the set of nodes be $V = NE \cup T$. Each node can be assigned occurrence statistics of the corresponding entity or term, such as document or sentence positions. To model *co*occurrence data, on the other hand, we utilize edge attributes in the network model.

Network Edges. Edges $E \subset V \times V$ describe the cooccurrence of entities and terms in documents, requiring that at least one node of an edge $e = (v_1, v_2)$

corresponds to an entity, i.e., $v_1 \in NE$ or $v_2 \in NE$. The cooccurrence of two respective words can be limited by a maximum sentence distance *sdist*. For $sdist = 0$, only cooccurrences in the same sentence are considered, for $sdist = 1$ the sentences directly before or after a given sentence, and so on.

Our model is based on three cooccurrence statistics, namely (1) the number of word cooccurrences, (2) the publication dates of the articles in which they cooccur, and (3) the textual distances at which the words cooccur. This scheme allows us to easily integrate new documents, eventually providing a basis for exploring the evolution of topics, as outlined in Sect. 4. To include the above features, each edge has an associated list of $\langle d, t, \delta \rangle$ tuples that encode the document d, timestamp t and the smallest mention distance δ (counted in sentences). For the first cooccurrence of two words v_1 and v_2 in a document d, we add the corresponding edge to the network, as well as the pair $\langle d, t(d), \delta(v_1, v_2) \rangle$ to that edge's list. If the same words cooccur again in the same document, we simply update the distance if necessary. If a cooccurrence of the two words is found in a new document, a new tuple is added to the list. In a sense, these lists represent a time series of word cooccurrences that support subsequent explorations, and enable efficient updates of the network representation. Similarly, we store lists of tuples $\langle d, t \rangle$ as node attributes for individual word mentions.

The resulting network serves as a model for a timestamped document collection, which is represented as collocations of words. In the following, we discuss how substructures in the network can be associated with topics in the documents.

4 Network-Based Topic Exploration

How are topics reflected in an implicit network? In the following, we argue that the core of topics is formed by edges between frequently cooccurring nodes, and that topics can be grown around such edges in a well-defined manner. We propose two growth approaches that specifically allow for an interactive exploration of topics. Finally, we discuss evolving topics based on the temporal edge labels.

4.1 Edge Weighting

Given the structure of news events, it appears to be a reasonable conjecture that a high cooccurrence frequency of two entities is indicative of a topic in news. For example, interactions between politicians, parties, countries, companies, or other actors and locations all involve more than one entity, which supports a faceted exploration of topics. Extracting and linking such named entities is an established task, so the question that remains is how such important edges can be identified and filtered from spurious connections in the network. For a naive approach, assume an edge $e = (v_i, v_j)$ with tuple list $L(e) = \langle (d_1, t_1, \delta_1), \ldots, (d_k, t_k, \delta_k) \rangle$. Let $w(e) = |L(e)|$ be the weight of that edge, that is, we use the length of the list to reflect the overall cooccurrence frequency. Clearly, an edge with a higher weight is more likely to be at the center of an important topic than an edge with a lower weight. Based on this intuition, we

introduce a weight for the edges of the graph G_D that supports such a filtering and includes both the overall and the temporal frequency of joint mentions, as well as the cooccurrence distances.

For an edge $e = (v_1, v_2)$, let $D(e) = \{d \mid (d, \cdot, \cdot) \in L(e)\}$ denote the set of documents in which v_1 and v_2 cooccur. Similarly, let $T(e) = \{t \mid (\cdot, t, \cdot) \in L(e)\}$ denote the set of timestamps at which both words that correspond to v_1 and v_2 occur jointly in a document. Let $D(v)$ and $T(v)$ be defined analogously for single nodes v. Finally, let $\Delta(e) = \langle \delta_1, \ldots, \delta_{|L(e)|} \rangle$ be the sequence of minimum distances at which the words that correspond to the two nodes of edge e occur in documents. We then obtain a combined weight for edge $e = (v_1, v_2)$ as

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

Intuitively, the measure represents the harmonic mean of three individual components, namely the number of joint versus individual mentions, the temporal coverage density, and an exponentially decaying weight by mention distance [18]. The resulting weight is normalized such that $\omega \in [0, 1]$.

Using the weights computed this way allows a pruning of low frequency edges and the detection of important entity connections, from which we grow and explore topics in the following. Since the components of the edge weights can be computed during network construction (or network updates with new documents), no additional post-processing costs occur during topic exploration.

4.2 Topic Construction and Edge Growth

Assuming an ordering of edges in G_D by weight, the top-ranked edges correlate to topic *seeds* as described above. Thus, one can select the top-ranked k edges for some value of k and treat them as seeds around which the topics are grown. Note that some seed edges may share nodes, an aspect that we discuss in Sect. 4.3. To grow topic substructures around the selected edges, we introduce two types of growth patterns, *triangular growth* and *external node growth*.

Triangular Growth. Given an edge $e = (v_1, v_2)$ between entities v_1 and v_2 along with a network substructure that only contains e, v_1, and v_2, this initial substructure can be grown by adding neighbours of both entities. Formally, let $N(v_1)$ and $N(v_2)$ denote the neighbours of nodes v_1 and v_2 respectively, then $N(v_1) \cap N(v_2)$ is the set of all nodes in G_D that share v_1 and v_2 as neighbours. To rank nodes in this potentially very large set, we utilize a scoring function on the edge weights. Specifically, let $s : V \to \mathbb{R}$ such that $s(x) = \min\{\omega(x, v_1), \omega(x, v_2)\}$. Obviously, nodes with a higher score cooccur more often and more consistently with both entities of the seed edge. Ranking nodes in the shared neighbourhood according to s thus allows us to select the most related terms to the topic that is represented by the seed edge. It is then a simple matter of adding any number of such nodes (along with the edges connecting them to v_1 and v_2) to the network substructure to incrementally grow the topic. Since all new nodes can be ranked

according to s, we obtain a relevance score for nodes in relation to the given edge e. In addition to the two seed words v_1 and v_2, a topic can thus be viewed as a list of ranked words that are added to the initial two words based on their cooccurrence patterns, much like a classic topic model. However, this growth strategy also results in a descriptive network substructure as illustrated in Fig. 1, where l such triangles are added to the seed edge e.

Based on the process described above, the incremental addition of words to the seed edge clearly supports different aspects of topic and cooccurrence exploration. First, instead of adding terms as described above, it is equally viable to select entities or even specific types of entities in the shared neighbourhood. For example, if v_1 and v_2 are both persons, one could restrict the growth process to add only other persons. Second, one should keep in mind that, depending on the realization of the network, the implicit network can be used as an inverted index [16], thus allowing the user to inspect articles and sentences in which two words cooccur during the incremental construction and exploration of the network substructures.

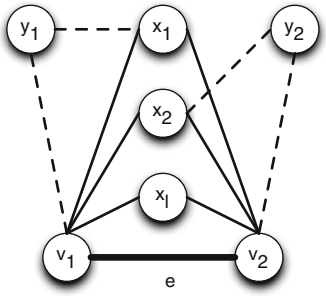

Fig. 1. Edge growth approach for seed edge e with words v_1 and v_2. Nodes $x_1, \ldots, x_l$ denote words added during triangular growth. Nodes y_1 and y_2 are (optional) additions in a subsequent external growth phase.

External Node Growth. While the construction of edge triangles and word rankings is a key component in extracting and exploring a topic in the classical sense, the node set determined in this way can also be explored based on further expansion techniques. As a topic substructure grows, the seed edge and its incident nodes are not the only available attachment points for further edges and nodes. Instead, one could also add further nodes that are connected only to one of the initial words v_1 or v_2, but also to some of the other nodes that were added in subsequent triangles. The external node growth process is also illustrated in Fig. 1, where nodes y_1 and y_2 connect to the substructure by dotted lines. While this attachment step has no analogy in classic topic extraction, it introduces additional degrees of freedom in an interactive exploration of topics.

4.3 Topic Fusion and Overlap

In the network-based extraction of topics, the substructure that is grown from the top-ranked seed edge is self-contained. However, this is not necessarily the case for substructures grown from subsequent edges in the list of top-ranked entity edges. Assume an edge $e' \neq e$ with $\omega(e') < \omega(e)$ from the list of seed edges. While new nodes are added to the substructure around e', it is possible that an edge is added that is incident to a node of the previously extracted substructure grown from e. In practice, this overlap between two topics may occur for terms

or entities. In the most extreme case, even seed edges may overlap in one of their entity nodes, leading to the fusion of two topics. In classic topic models, the same word may belong to different topics with a high probability, which is analogous to partially overlapping topics in our model, where the same node can be part of different topics. In fact, we argue that topics should overlap for entities in news that participate in multiple topics, and in Sect. 5 we show how network visualization highlights such overlapping substructures during topic exploration.

4.4 Evolving Topics

An important aspect of topic modeling is the evolution of topics over time (e.g., [3,10]). In such a setting, visualizations turn out to be especially helpful in highlighting changes in a topic's relevance over time or how it compares to other topics. Our framework directly supports the exploration of temporal topic characteristics due to the timestamp information contained in the edge labels.

Key to the exploration of evolving topics based on an implicit network is the *network projection*. In general, a network projection filters nodes and edges that do not satisfy certain user-specified conditions. For example, a given network could be projected to only cooccurrences of entities, i.e., all term nodes and incident edges are removed. For studying the evolution of topics, such conditions concern the timestamps associated with the cooccurrence information of edges. In principle, given a collection of news articles spanning a time interval from t_{min} to t_{max}, the interval can be partitioned to construct corresponding networks. For example, the evolution of news topics over multiple weeks or months can be considered by focusing on multiple networks constructed for these intervals.

For a seed edge in such an interval, one can then directly compare respective sub-networks from different time intervals side-by-side, thus highlighting what nodes have gained or lost relevance for that topic in a given interval with respect to neighbouring intervals. There are numerous visualization metaphors one can consider in this setting, all of them relying on the visualization of sub-networks. Key to all these approaches is the representation of topics in the (entity-centric) context of a network structure that explicitly represents the implicit word relations in the documents. In the following, we give an example of such a process for exploring the evolution of topics over time in a document collection.

5 Evaluation

We give a description of the news article data used in our exploration of topics, before presenting the results for topic networks and a comparison to LDA topics.

5.1 News Article Data

To demonstrate the advantages of a network-based approach and investigate the evolution of topics, we focus on the extraction of topics from news articles that provide both a temporal component and a large scale. Since we are unaware of

existing data sets that are sufficiently large and annotated for named entities (or even topics), we collect the articles from the RSS feeds of a variety of international outlets. For reproducibility, we make the list of article URLs available[1].

Data Collection. We collect all articles from 14 English-speaking news outlets located in the U.S. (CNN, LA Times, NY Times, USA Today, CBS News, The Washington Post, IBTimes), Great Britain (BBC, The Independent, Reuters, SkyNews, The Telegraph, The Guardian), and Australia (Sidney Morning Herald). We collect all their feeds that are related to political news from June 1, 2016, to November 30, 2016. After removing articles that have less than 200 or over 20,000 characters or more than 100 identified entities per article (e.g., lists of real estate sales), we obtain 127,485 articles with 5.4M sentences.

Data Preparation. We use manually created extraction rules for each news outlet to strip the HTML code and cleanly extract the text, before performing sentence splitting, tokenization, entity recognition, entity linking, entity classification, and stemming. For the recognition and classification of named entities, we use the Ambiverse API [2], which disambiguates entity mentions to Wikidata IDs. To classify named entities into persons, locations, and organizations, we map Wikidata entities to YAGO3 entities and classify them according to the YAGO hierarchy since a classification in Wikidata is problematic [17]. We use the class `wordnet_person_100007846` for persons, `wordnet_social_group_107950920` for organizations and `yagoGeoEntity` for locations. For sentence splitting and part-of-speech tagging, we use the Stanford POS tagger [19].

Network Construction. To construct the network, we use a modified version of the LOAD implicit network extraction code [18] that we adapted to utilize disambiguated entities and add document timestamps and outlet identifiers to edges. Terms are stemmed with the Snowball Porter stemmer[3]. We set the window size for the extraction of entity cooccurrences to $sdist = 5$. The resulting network has 27.7 k locations, 72.0 k actors, 19.6 k organizations, and 329 k terms, which are connected by 10.6 M labelled edges. To generate edge weights for the resulting network, we use the weighting scheme ω described in Sect. 4.1.

5.2 Entity-Centric Extraction of Topics

As a first step of our exploration, we consider the extraction of traditional topics as lists of words with importance weights. Based on the underlying assumption that topics are focussed on entities, we first obtain a ranking by weight of all edges in the network that connect two entities. Thus, we utilize a global ranking of edges to identify relevant seeds for topics, which stands in contrast to the local entity-centric approaches that have been used on implicit networks so far [16,18]. The top-ranked edges are then considered to form the seeds of

[1] The URLs of articles in our data, the extracted implicit network, and our program code are available at https://dbs.ifi.uni-heidelberg.de/resources/nwtopics/.
[2] https://www.ambiverse.com/.
[3] http://snowballstem.org/.

Table 1. Traditional topics as ranked lists of terms, extracted for the four top-ranked edges in the network generated from the subset of NewYork Times articles. For each edge, the two incident entities and their Wikidata identifiers are given.

Beirut - Lebanon		Russia - Moscow		Russia - Putin		Trump - Obama	
Q3820 - Q822		Q159 - Q649		Q159 - Q7747		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

topics. Subsequently, each such edge is grown to a topic description by adding neighbouring terms that are connected to both entities as described in Sect. 4.2. The resulting ranking can be utilized as a topic. Since only the local neighbourhood of the two entities is considered once the global ranking is obtained, this process is extremely efficient, can be parallelized by edge and computed at query time to obtain, expand, or reduce an arbitrary number of topics interactively.

As an example, we show the topics that are induced by the four highest ranked entity edges in the subset of NewYork Times articles in Table 1 (topics are ranked from left to right by the value of ω of the seed edge). We find that the topics are overall descriptive and can be interpreted within the context of news in 2016. With regard to location mentions, the example shows well a prevalent bias in news articles, which often include the location of the correspondent or news agency at the start of the article (i.e., articles about the war in Syria are often not reported from Syria itself but neighbouring Lebanon). However, even when seed edges overlap in a common entity, the resulting topics are still descriptive and nuanced, as the example of Russia shows, which is associated with aspects of Russian politics and the involvement of Putin. If topics that are focused on such synonyms (e.g., the mention of a capital instead of a country) are not of interest, filtering edges by entity type is easily possible.

Much like traditional topic models, the topics and topic qualities that we find vary strongly by news outlet, as we show in Sect. 5.4. Since the topics of each edge are independent, it is easier to discard unwanted topics than it would be for traditional topic models with interdependent topics. Overall, we find that we can replicate list-based topics with an edge-centric approach. However, a representation of topics as lists of terms is needlessly abstract and the network supports the extraction of visually descriptive topics, as we show in the following.

5.3 Topic Network Extraction and Exploration

To fully utilize the network representation, we extract topic substructures. Instead of lists of terms from nodes that surround seed edges, we extract the nodes themselves to continually grow a descriptive network structure. We proceed in the same way as described for traditional topics above by extracting a ranked list of entity-centric edges and selecting the top-ranked edges. For each edge, we then include a number of terms that are adjacent to both entities in the network. The number of adjacent nodes can be selected arbitrarily, but a value of around three term nodes per edge tends to result in a visually interpretable network. By projecting the network according to the publication date of the corresponding article, we can introduce a temporal dimension and investigate the evolution of topic networks over time or even dynamically for selected intervals.

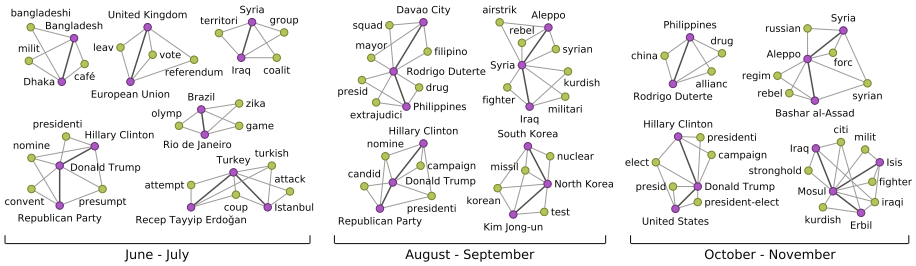

Fig. 2. Topic substructures of entities (purple) and terms (green) for the article subset of CNN. Shown are the 8 highest ranked edges and the 3 most relevant connected terms. The data is divided into three segments of two months to highlight topic evolution. (Color figure online)

In Fig. 2, we show temporal snapshots for the CNN article subnetwork. The results for other news outlets are similar (albeit with a regional or political bias). While we still find the same descriptive terms in the graph representation as we do in the case of ranked term lists, we also observe the additional structure of the underlying network. Unlike term lists, which represent isolated topics, the overlaps of edges show topic relations directly. In fact, we observe entity-entity subgraph structures that emerge from the top-ranked edges and lend further support to their topics. For example, the topics of Trump, Clinton, and the U.S. are clearly related. On the temporal axis, we find that the topics correlate well with political events. For example, the Brexit topic disappears after the referendum in June 2016 (of course, it is more pronounced in British outlets), while several war-related topics shift focus to follow ongoing campaign locations, and the US election topic is expectedly stable. Overall, the network representation adds a structure to the visualization that is easily recognizable and explorable.

5.4 Comparison to LDA Topics

It is well known that topics are subjective and a strict evaluation is difficult, especially for an exploratory approach. To relate network topics to traditional topic models, we compare their list-of-term representation to LDA topics [4]. We extract topics for each news outlet from the network as described in Sect. 5.2. For LDA topics, we group the news articles by outlet, prepare the plain text with Tidytext[4], and generate topics with the R implementation of LDA[5]. As a metric for the comparison, we compute the coverage to capture how well each of the topics that are produced by one approach are reflected in topics produced by the other approach. Formally, for two sets of topics T and U of size $k = |T| = |U|$, we let

$$coverage(T, U) := \frac{1}{k} \sum_{t \in T} \max_{u \in U} \{jaccard(u, t)\}.$$

Note that the coverage as defined above is not symmetric. Since entities are a major component of network topics, we add the tokenized labels of seed nodes to the term lists before selecting the top-ranked terms of a network topic.

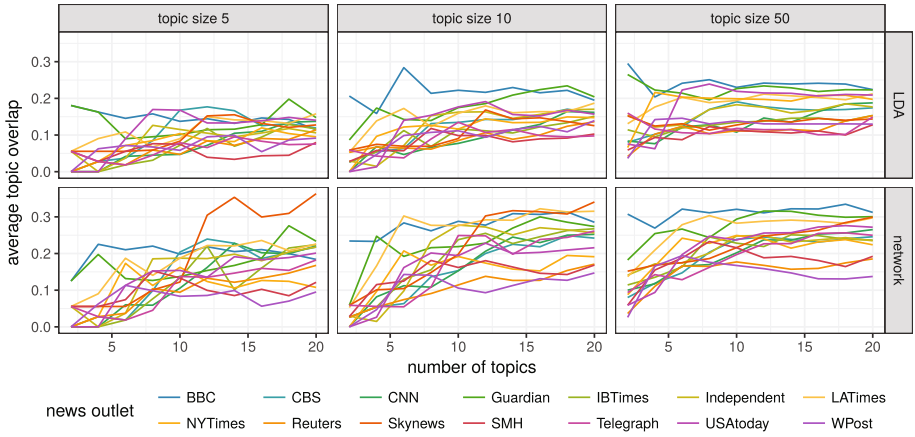

Fig. 3. Comparison of coverage between network topics and LDA topics. Topic size denotes the number of words per topic. Results are shown for the coverage of LDA topics in network topics (top row) and network topics in LDA topics (bottom row).

Figure 3 shows the comparison results for all 14 news outlets. The overall coverage increases with the number of topics and with the number of terms per topic. However, the coverage of LDA topics in network topics is much worse than the other way around. Combined with the increasing coverage for larger numbers of topics, this indicates the more narrowly focussed nature of network topics. The

[4] https://cran.r-project.org/web/packages/tidytext/.
[5] https://cran.r-project.org/web/packages/topicmodels/.

coverage of LDA topics in network topics increases with the number of topics since this narrows the scope of LDA topics. Overall, we find that network topics seem to be well reflected in LDA topics once the number of extracted topics for LDA is large enough to distinguish between the multitude of news topics. However, for a number of topics beyond 20, the runtime of LDA becomes an serious issue on the larger news outlets since the number of topics is a multiplicative factor in the runtime complexity of LDA, while it is an addend in the complexity of network topics. Furthermore, the number of network topics can be dynamically adjusted during the exploration phase and supports settings where the repeated extraction of traditional topics is too compute intensive.

6 Conclusions and Ongoing Work

The detection of topics in a (dynamic) collection of documents is a central task in corpus analysis, in particular for document classification and clustering. While there exist many approaches for topic modeling, they often fall short in terms of intuitive, interactive, and efficient topic exploration methods. In this paper, we presented a network-based, entity-centric framework for topic exploration in collections of news articles that addresses these needs and provides a novel and intuitive view on topics as network substructures instead of ranked lists of words. Based on word cooccurrences, we showed that a network can be efficiently constructed and updated from a corpus and that such a network supports the exploration of topics and their relationships. In our ongoing work, besides studying the applicability of our approach to other corpora beyond news articles, we are investigating network approaches to further topic analysis needs. In particular, we are developing easy-to-use interfaces for network-based topic exploration.

Acknowledgements. We would like to thank the Ambiverse Ambinauts for kindly providing access to their named entity linking and disambiguation API.

References

1. Blanco, R., Lioma, C.: Graph-based term weighting for information retrieval. Inf. Retr. **15**(1), 54–92 (2012)
2. Blei, D.M.: Probabilistic topic models. Commun. ACM **55**(4), 77–84 (2012)
3. Blei, D.M., Lafferty, J.D.: Dynamic topic models. In: ICML (2006)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
5. Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J.L., Blei, D.M.: Reading tea leaves: how humans interpret topic models. In: NIPS (2009)
6. Evert, S.: The statistics of word cooccurrences: word pairs and collocations. Ph.D. thesis, University of Stuttgart, Germany (2005)
7. Gretarsson, B., O'Donovan, J., Bostandjiev, S., Höllerer, T., Asuncion, A., Newman, D., Smyth, P.: TopicNets: visual analysis of large text corpora with topic modeling. ACM Trans. Intell. Syst. Technol. **3**(2), 23:1–23:26 (2012)

8. Gries, S.T.: 50-something years of work on collocations. Int. J. Corpus Linguist. **18**(1), 137–166 (2013)
9. Han, X., Sun, L.: An entity-topic model for entity linking. In: EMNLP (2012)
10. Hong, L., Yin, D., Guo, J., Davison, B.D.: Tracking trends: incorporating term volume into temporal topic models. In: KDD (2011)
11. Hu, Y., Boyd-Graber, J., Satinoff, B., Smith, A.: Interactive topic modeling. Mach. Learn. **95**(3), 423–469 (2014)
12. Newman, D., Chemudugunta, C., Smyth, P.: Statistical entity-topic models. In: KDD (2006)
13. Rousseau, F., Vazirgiannis, M.: Graph-of-word and TW-IDF: new approach to ad hoc IR. In: CIKM (2013)
14. Sarma, A.D., Jain, A., Yu, C.: Dynamic relationship and event discovery. In: WSDM (2011)
15. Shi, B., Lam, W., Jameel, S., Schockaert, S., Lai, K.P.: Jointly learning word embeddings and latent topics. In: SIGIR (2017)
16. Spitz, A., Almasian, S., Gertz, M.: EVELIN: exploration of event and entity links in implicit networks. In: WWW Companion (2017)
17. Spitz, A., Dixit, V., Richter, L., Gertz, M., Geiss, J.: State of the union: a data consumer's perspective on Wikidata and its properties for the classification and resolution of entities. In: Wikipedia Workshop at ICWSM (2016)
18. Spitz, A., Gertz, M.: Terms over LOAD: leveraging named entities for cross-document extraction and summarization of events. In: SIGIR (2016)
19. Toutanova, K., Klein, D., Manning, C.D., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: HLT-NAACL (2003)
20. Zuo, Y., Zhao, J., Xu, K.: Word network topic model: a simple but general solution for short and imbalanced texts. Knowl. Inf. Syst. **48**(2), 379–398 (2016)