TeMoTopic: Temporal Mosaic Visualization of Topic Distribution, Keywords, and Context

Shane Sheehan, Saturnino Luz
University of Edinburgh
United Kingdom
shane.sheehan@ed.ac.uk
s.luz@ed.ac.uk

Masood Masoodian
Aalto University
Finland
masood.masoodian@aalto.fi

Abstract

In this paper we present *TeMoTopic*, a visualization component for temporal exploration of topics in text corpora. *TeMoTopic* uses the temporal mosaic metaphor to present topics as a timeline of stacked bars along with related keywords for each topic. The visualization serves as an overview of the temporal distribution of topics, along with the keyword contents of the topics, which collectively support detail-on-demand interactions with the source text of the corpora. Through these interactions and the use of keyword highlighting, the content related to each topic and its change over time can be explored.

1 Introduction

Many text corpora, such as news articles, are temporal in nature, with the individual documents distributed across a span of time. As the size and availability of text corpora have continued to increase in recent years, effective analysis of the content of corpora has become challenging. Taking the temporal nature of most corpora into account when analysing the text makes it more difficult to describe the corpora and to interpret intuitively the results of analysis.

Topic modeling techniques, such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003), have been used to automatically generate topic groups in text corpora. These topics can help in understanding the contents of a corpus by using keywords and topic association probabilities generated by the topic modelling technique. However, interpreting the results of the techniques is not always easy, and the results can seem counter-intuitive when looking only at the weighted keyword lists. Therefore, visualization techniques have been used extensively to help with the interpretation of the large number of topics generated by these models. The same is true of temporal topic modeling techniques, such as Dynamic Topic Modeling (Blei and Lafferty, 2006), which require additional visualization techniques to aid intuitive understanding of the temporal segmentation of the topics and their related keywords.

In this paper, we propose *TeMoTopic* as a contribution to the collection of visualization techniques for exploring the temporal distribution of topics in text corpora through the use of temporal mosaics. *TeMoTopic* adopts a space-filling approach to show topic distribution over time, and presents keywords related to each topic at the overview level of the visualization. The visualization is interactive and, in contrast to many other techniques, enables direct investigation of the source documents associated with individual topics and keywords. This allows the user to get a general sense of the meaning of a topic through its associated keywords, as well as providing the ability to dive into the details of the related documents.

2 Related Work

2.1 Temporal Topic Visualization

Topic visualization systems are an active research area, with a variety of approaches for visualizing different aspects of topic model outputs, topic hierarchies, and topic evolution. In this paper, we only focus on related work in the area of temporal topic evolution and topic visualization of text corpora. While some methods address the temporal structuring of topics in short texts in the context of meetings and dialogues (Luz and Masoodian, 2005; Sheehan et al., 2019), in recent years, visualization of temporal topic evolution for larger text collections has been based on flow diagrams. An early example of such an approach is *ThemeRiver* (Havre et al., 2002), with later additions such as *TextFlow* (Cui et al., 2011), *TopicFlow* (Malik et al., 2013), *ThemeDelta* (Gad et al., 2015) and *RoseRiver* (Cui et al., 2014).

While *TeMoTopic* and flow-based temporal topic visualizations are similar, we expect they could

Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation, pages 56–61

Task	Description
Visualize Topics	Visualize topic in terms of extracted keywords
Overview of Document - Topic Relations	View documents related to a topic
Remove Topics from the visualization	Topic removal from overview
Filtering Documents	View a subset of documents for a topic
Perform Set Operations	Enable exclusion/inclusion of documents in the corpus
Show and Cluster Similar Topics	Enable identification of similar topics
Perform Cluster Operations	Enable grouping of similar topics
Annotating Topics	Allow for labelling of the topics
Visualize Topic Change	View topic distribution and keywords over time

Table 1: Visualization tasks for topic model exploration.

form complementary components used in model assessment tools that are used to evaluate model quality. Flow diagrams are, for instance, useful for getting a high-level overview of many topics across long spans of text. *TeMoTopic*, on the other hand, aims to provide support for detailed viewing of a subset of topics and shorter timeslices, which are not possible in a flow diagram. As such, we envisage that other existing visualization tools which include a flow diagram component – such as *LDAExplore* (Ganesan et al., 2015), *VISTopic* (Yang et al., 2017), *ParallelTopics* (Dou et al., 2011) and *TIARA* (Wei et al., 2010) – could be further expanded to include a temporal mosaic visualization, in the style of *TeMoTopic*. The largest benefit to this integration would come from enabling intuitive interactive filtering of the source documents based on the temporal topic and keyword distribution.

2.2 Topic Visualization Tasks

The design of a visualization tool should clearly be motivated by concrete tasks relevant to the end-users of the intended tool. Munzner's *nested model for visualization design and validation* (Munzner, 2009) describes steps that can be taken to mitigate threats to the validity of a visualization design. The first of the four levels of this design model is the characterization of domain specific tasks which should be supported by the visual encoding.

Ganesan et al. (2015) identify key tasks, in the design description of *LDAExplore*, which should be supported by visualizations that aim to help users explore the results of Latent Dirichlet Allocation (LDA). Since LDA is one of the most commonly used topic modelling techniques for text corpora, these key tasks could be generalized to other techniques where a corpus is also split into topics, and keywords associated with those topics are extracted.

In addition, Ganesan et al. (2015) argue that the results of LDA can be counter-intuitive, and that the ability to explore and interact with the document set should make the topic and word distributions more intuitive and insightful. Table 1 shows the eight tasks identified by Ganesan et al. (2015), as well as one additional task which we consider to be important for visualizing temporal topics. The table also includes a brief description of the tasks which are fully described by Ganesan et al. (2015).

Theses tasks describe a need for topic overview with document detail available on-demand, this follows the well-known visual information seeking mantra proposed by Shneiderman (1996). Interactions around viewing, filtering, removing, and combining topics and documents should also be supported. Finally, we include an additional task for visualizing topic changes over time. This modifies the *Visualize Topics* task, such that the change in distribution and keywords across is available to explore.

3 TeMoTopic: Temporal Mosaic Topic visualization

Figure 1 shows the *TeMoTopic* visualization tool. It consists of two juxtaposed views (Javed and Elmqvist, 2012): the temporal mosaic (left), and the document view (right). The design of the temporal mosaic is based on a visualization proposed by Luz and Masoodian (2007), and further expanded in our previous temporal mosaic visualizations *TeMoCo* visualization (Sheehan et al., 2019) and *TeMoCo-Doc* visualization (Sheehan et al., 2020), which have been used to link transcripts of meetings to document reports in a medical context.

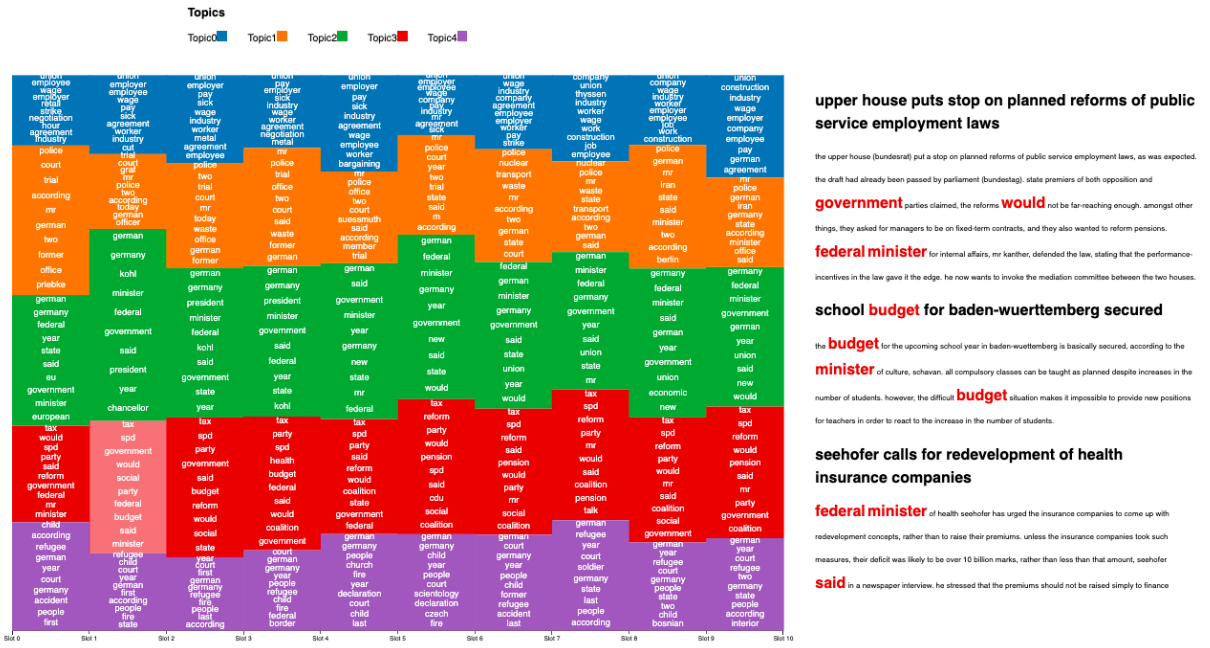

Figure 1: *TeMoTopic* visualization, showing the temporal mosaic view (left) and the document view (right), showing the selected keywords for the red topic in the second timeslice (red tile on the bottom left).

Figure 2: *TeMoTopic* visualization, showing the temporal mosaic view (left) and the filtered document view (right), with the word "german" selected from a temporal topic timeslice (orange tile on the top left).

3.1 Prototype

The temporal mosaic encoding was designed using Mackinlay's ranking (Mackinlay, 1986) of visual variables (Bertin, 1983), such that the visualization uses a perceptually efficient static encoding of the key data attributes. Horizontal position is used to emphasize the temporal order of the topics, and topic distribution per timeslice is encoded using vertical length. Each tile in the mosaic represents a single combination of topic and timeslice. The height of each tile represents its topic weight in that timeslice.

The top ten keywords which describe the associated temporal topic are placed within the tile, and can be scaled to encode the keyword topic probability, using area in a manner similar to keyword scaling in text visualizations such as word clouds (Viegas et al., 2009). Although the keywords are currently presented in order of descending topic probability, in future work alternative keyword presentation styles such as alphabetized lists and word clouds will be compared in terms of their effectiveness for comparison between the tiles. The categorical topics are encoded using color, allowing topics weights and keyword changes to be examined across the span of timeslices.

The mosaic visualization provides an overview of the topic distribution and associated keywords over time. However, as the number of topics and

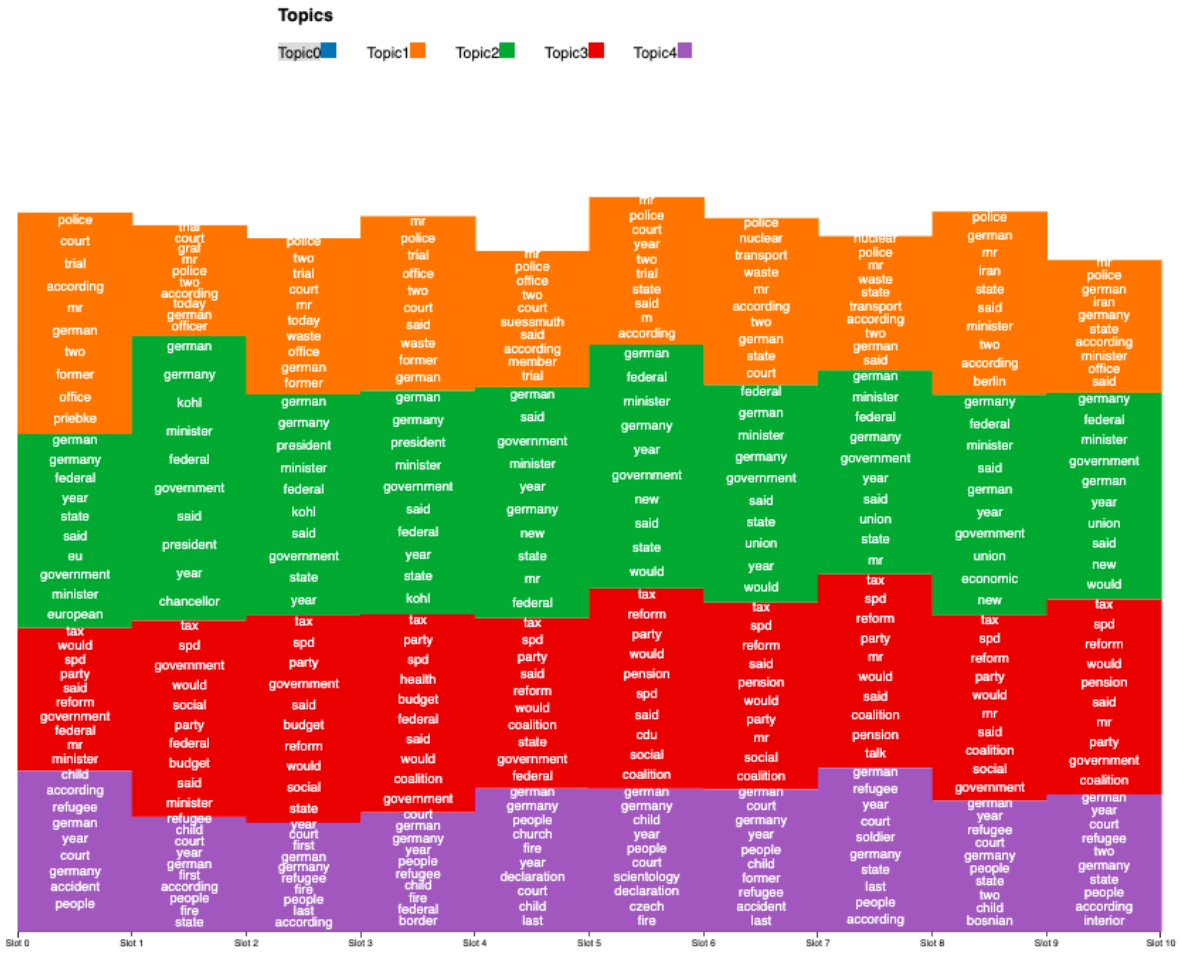

Figure 3: *TeMoTopic* filtered temporal mosaic view after the blue topic was selected for removal via clicking on the legend.

timeslices increase, if the visualization area is kept at a fixed size, the overview would become more abstract, cluttered, and difficult to examine for individual tiles and keywords. To maintain readability, the visualization can extend both horizontally and vertically to accommodate more topics and timeslices. The user can pan and zoom to get the detailed views of topics and keywords, or a higher-level view of the entire temporal topic space. The removal interaction is particularly useful when the number of topics is large, since filtering out topics that are not relevant to the current analysis allows for more of the detail to be presented on a single screen.

The temporal mosaic, as currently described, addresses two of the tasks from Table 1, namely *Visualize Topics* and *Visualize Topic Change*. To facilitate *Overview of Document - Topic Relations*, the document view (Figure 1, right) was created and linked, via click interactions, to the temporal mosaic (Figure 1, left). The document view is used to display the documents associated with a temporal topic tile. When a coloured tile is selected in the temporal mosaic, the related articles are presented in a scroll box and, the keywords from the topic tile are highlighted in the text. If keyword weights (or probabilities) are provided, the highlighted words are scaled accordingly. This dual combination of views and described interactions, support the user in investigating the meaning of a topic, and by investigating the differences between the topic timeslices, temporal document similarities and differences can be revealed.

Although it is useful to view the entirety of a topic, *Filtering Documents* is a task that was also identified as important to facilitate. One simple and intuitive way to do this with the temporal mosaic is by clicking on individual keywords rather than on the entire topic tile. This will cause the document view to display only documents from the related topic timeslice which contain the selected keywords, as shown in Figure 2. Selection from

multiple topics is also possible, and the keywords are highlighted in the related topic colour to differentiate between topics.

The final interaction supported by this version of *TeMoTopic* is the removal of topics from the temporal mosaic. To do this, a topic can be selected from the legend shown above the temporal mosaic (Figure 3, top). Alternatively right-clicking on a topic removes all the other topics except the selected one. In the example shown in Figure 3, the blue topic has been removed from the temporal mosaic. When topics are removed, the temporal mosaic no longer fills the entire vertical space of the visualization. This interaction is useful when dealing with a large number of topics of which only a few are of interest for the analysis.

3.2 Implementation

The visualization tool[1] is implemented as a single-page web application using the *D3.js* framework (Bostock et al., 2011) . It takes two JavaScript Object Notation (JSON) files as input: the first file contains topic, keyword, timeslice, weights, and associated filenames, and the second input file is simply a JSON structure containing the documents with filename used as the retrieval key. Sample Python scripts are provided for generating topics and keywords on the sample dataset and for preparing the visualization input files from the model output.

The current version of *TeMoTopic* was designed to be model agnostic, and can even be used for tasks unrelated to topic model exploration. For example, metadata attributes such as the source of the news articles or their author could be used in place of topics. Keywords could be extracted using any available technique, including simple frequency lists. The visualization could also be used for corpus comparison and even cross-lingual analysis using entire corpora as replacements for the topics.

However, in our implementation we make use of dynamic topic modelling (Blei and Lafferty, 2006) to identify temporal topics and keywords in a subset of the *de-news*[2] corpus of German-English parallel news. The dataset consists of transcribed German radio broadcasts which were manually translated into English. Between 1996 and 2000 volunteers selected and transcribed five to ten of these news broadcasts per day and added them to the dataset. In the examples of *TeMoTopic*, shown in Figures 2, 1 and 3, we selected a ten month span of the dataset and presented the four largest topics. The choice of time span and topic number was only for presentation and to exemplify the interface features. We did not attempt to choose a time period or number of topics based on prior knowledge of the news relevant at the time in Germany. We present our examples to describe the interface and interactions, rather than as an analysis of the dataset, and we choose to draw no conclusions about the dataset contents and topics.

4 Conclusions

While many other temporal visualization techniques, such as ThemeRiver (Havre et al., 2002), offer some of the functionality for temporal visualization of topics or visualization of content changes, they do not feature implicit linking between the visualization and the underlying content documents. We consider this to be the main contribution of *TeMoTopic* visualization and its distinguishing feature with regards to the state of the art. As such, determining the necessity and validity of this approach in the identified domain is an important step before further development of the visualization prototype. Future work will, therefore, include evaluating the usability of a future iteration of the system with domain experts in both news analysis and topic modelling.

Acknowledgments

The work of the first and second authors is supported by European Union's Horizon 2020 research and innovation programme under grant agreement No 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). The results of this paper reflects only the authors' view and the Commission is not responsible for any use that may be made of the information it contains.

References

Jacques Bertin. 1983. *Semiology of Graphics*. University of Wisconsin Press.

David M Blei and John D Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120.

[1]The software and working example are available at https://github.com/sfermoy/TeMoCo.

[2]http://homepages.inf.ed.ac.uk/pkoehn/publications/de-news/

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. 2011. D3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309.

W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, and X. Tong. 2011. Textflow: Towards better understanding of evolving topics in text. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2412–2421.

W. Cui, S. Liu, Z. Wu, and H. Wei. 2014. How hierarchical topics evolve in large text corpora. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2281–2290.

Wenwen Dou, Xiaoyu Wang, Remco Chang, and William Ribarsky. 2011. Paralleltopics: A probabilistic approach to exploring document collections. In *2011 IEEE conference on visual analytics science and technology (VAST)*, pages 231–240. IEEE.

Samah Gad, Waqas Javed, Sohaib Ghani, Niklas Elmqvist, Tom Ewing, Keith N Hampton, and Naren Ramakrishnan. 2015. Themedelta: Dynamic segmentations over temporal topic models. *IEEE transactions on visualization and computer graphics*, 21(5):672–685.

Ashwinkumar Ganesan, Kiante Brantley, Shimei Pan, and Jian Chen. 2015. Ldaexplore: Visualizing topic models generated using latent dirichlet allocation. *arXiv preprint arXiv:1507.06593*.

S. Havre, E. Hetzler, P. Whitney, and L. Nowell. 2002. Themeriver: visualizing thematic changes in large document collections. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):9–20.

Waqas Javed and Niklas Elmqvist. 2012. Exploring the design space of composite visualization. In *Pacific Visualization Symposium (PacificVis), 2012 IEEE*, pages 1–8.

Saturnino Luz and Masood Masoodian. 2005. A model for meeting content storage and retrieval. In *Proceedings of the 11th International Multimedia Modelling Conference*, MMM '05, pages 392–398.

Saturnino Luz and Masood Masoodian. 2007. Visualisation of parallel data streams with temporal mosaics. In *Proceedings of the 11th International Conference Information Visualization*, IV '07, pages 197–202.

Jock Mackinlay. 1986. Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics*, 5(2):110–141.

Sana Malik, Alison Smith, Timothy Hawes, Panagis Papadatos, Jianyu Li, Cody Dunne, and Ben Shneiderman. 2013. Topicflow: Visualizing topic alignment of twitter data over time. In *Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining*, pages 720–726.

T. Munzner. 2009. A nested model for visualization design and validation. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):921–928.

Shane Sheehan, Pierre Albert, Masood Masoodian, and Saturnino Luz. 2019. Temoco: A visualization tool for temporal analysis of multi-party dialogues in clinical settings. In *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 690–695. IEEE.

Shane Sheehan, Saturnino Luz, Pierre Albert, and Masood Masoodian. 2020. Temoco-doc: A visualization for supporting temporal and contextual analysis of dialogues and associated documents. In *Proceedings of the International Conference on Advanced Visual Interfaces*, AVI '20, New York, NY, USA. Association for Computing Machinery.

Ben Shneiderman. 1996. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings the IEEE Symposium on Visual Languages*, pages 336–343.

Fernanda B. Viegas, Martin Wattenberg, and Jonathan Feinberg. 2009. Participatory visualization with wordle. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1137–1144.

Furu Wei, Shixia Liu, Yangqiu Song, Shimei Pan, Michelle X Zhou, Weihong Qian, Lei Shi, Li Tan, and Qiang Zhang. 2010. Tiara: a visual exploratory text analytic system. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 153–162.

Yi Yang, Quanming Yao, and Huamin Qu. 2017. Vistopic: A visual analytics system for making sense of large document collections using hierarchical topic modeling. *Visual Informatics*, 1(1):40–47.