Detecting policy fields in German parliamentary materials with Heterogeneous Information Networks and Node Embeddings

Alexander Brand [1] Wolf J. Schünemann [1] Tim König [1]
Tanja Preböck [2]

[1]Hildesheim University

[2]University of Bamberg

1st Workshop on Computational Linguistics for Political Text Analysis

Mapping and Retrieval for German Policy Debates

- Part of the EPINetz Project (Exploration politischer Informationsnetzwerke; https://epinetz.de/)
- Dynamic and permable ontology of policy debates
- Exploration of new ways to integrate multiple units of information (e.g. entities and interactions)

Theoretical foundations and the state of research

- Movement from studying traditional preoccupation with stateled political steering to more flexible and heterogeneous governance constellations; from institutional to ideational (Sabatier, 1988)
- Integration of network-based approaches (Freeman, 2004; Heclo, 1978; Heinz, Laumann, Salisbury, & Nelson, 1990; Kenis & Schneider, 1991; Laumann & Knoke, 1987)
- More and more computational methods (Lazer & Wojcik, 2018)

Theoretical foundations and the state of research

- Interpretative discourse research for informed entity and relation selection (Fischer, 2003; Hajer, 2002; Yanow, 2009)
- Even more computational methods, like word embeddings and node embeddings for social science research (Rheault & Cochrane, 2020; Rodriguez & Spirling, 2021; Won & Fernandes, 2021)
- Extending established combinations of discourse and policy research (Leifeld, 2018)

Methodological Considerations

We take a look at:

- Heterogenous Information Networks
- Metapath-based Embeddings
- Clustering (in the example)
- Near Neighbours (in the example)

Heterogeneous Information Networks

- Definition: A Heterogeneous Information Network (HIN) is a graph, which can consist of multiple node and/or edge types (Sun & Han, 2013).
- We formalise a HIN as $G = (V, E)$ with a node type mapping of $\phi : V \to A$ (node types). We construct meta-paths P which describe paths on the graph of the network scheme $T_G = (A, R)$ and have the form of $A_1 \xrightarrow{R}_1 A_2 \xrightarrow{R}_2 \ldots \xrightarrow{R}_l A_{l+1}$ (Sun & Han, 2013)

Metapath-based Embeddings

- Similar to Word Embeddings (Grover & Leskovec, 2016; Rheault & Cochrane, 2020)
- Better representations than node2vec (Dong, Chawla, & Swami, 2017)
- Are able to map complex relationships in networks (Dong et al., 2017)

Dataset

- "every single word" dataset (Remschel & Kroeber, 2020)
- Contains all written communication (reports, petitions, etc.) published by the German Bundestag between 1949 and 2017
- Restriction of the the sample to the 10th election period (1983-1987; first period with Greens ("Die Grünen"))
- $N = 6534$ documents

Analysis

- Network with Types $\mathbb{A}$ with $V = 22763, E = 792127$
  - Committee (Com) (e.g. parliamentary committee)
  - Fraktion/Bundes* (F/B*) (parliamentary faction, federal institution)
  - Keyword (extraction via tf-idf (Aizawa, 2003))
  - Named Entity (matched via the GermaParl Corpus (Falk & Meisinger, 2020))
  - Document
- Centred around documents
- Connections symbolize inclusion/authorship

Analysis

- Construction of Metapaths $\mathbb{P}$ of
  - $Document \longrightarrow Com \longrightarrow Document$
  - $Document \longrightarrow F/B^* \longrightarrow Document$
  - $Document \longrightarrow Keyword \longrightarrow Document$
  - $Document \longrightarrow Entity \longrightarrow Document$
- Embedding with random walker with steps $= 100$ (Stellargraph, 2020)
- Dimension Reduction via T-SNE and Clustering via dbscan (Ester, Kriegel, Sander, & Xu, 1996; Stellargraph, 2020; van der Maaten & Hinton, 2008) with minimum cluster size $=$ 50 Nodes, checked via average silhouette width $\overline{S_i}$.
- Qualitative coding according to the policy agenda project for the top 10 documents with the highest degree in the clusters (Jennings & Bevan, 2012; Sarshar & Roychowdhury, 2005)

Results

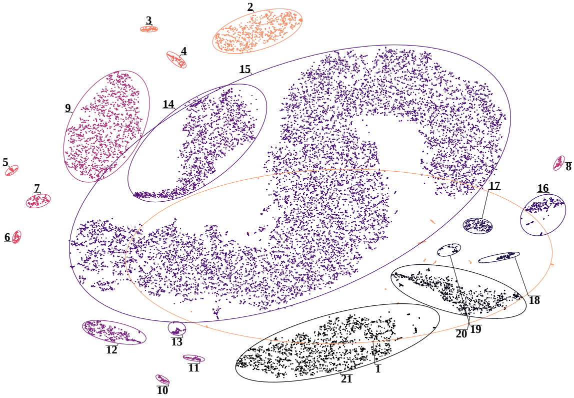
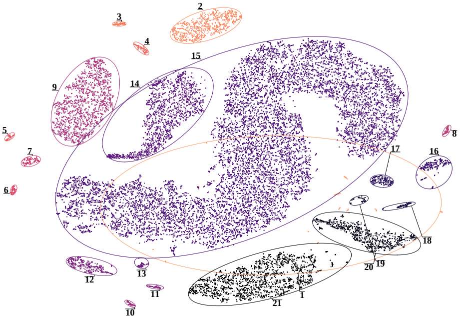


Figure: Visualisation of the Node Embeddings via Metapath2Vec

Results



Cluster	N	$\overline{S_i}$
1 LCFI	184	-0.61
2 Social Welfare	853	0.59
3 Agriculture	75	0.89
4 Environment	105	0.84
5 BFD; Social Welfare	68	0.9
6 Social Welfare	77	0.9
7 EU Affairs	132	0.82
8 Labour and Employment	77	0.92
9 Government Operations	1736	0.34
10 Environment; BFD; Mixed	67	0.89
11 CMIC	76	0.85
12 Macroeconomics	355	0.66
13 Government Operations	72	0.86
14 Infrastructure	1854	0.37
15 Government Operations	12846	-0.33
16 Mixed	292	0.61
17 CMIC	195	0.74
18 Macroeconomics	103	0.75
19 Government Operations	66	0.8
20 Energy	902	0.27
21 Government Operations	2007	0.36

Table 1: Average silhouette width by cluster. Higher values indicate better separation. BFD = Banking, Finance, and Domestic Commerce; CMIC = Civil Rights, Minority Issues, Immigration and Civil Liberties; LCFI = Law, Crime, and Family Issues

Figure: Visualisation of the Node Embeddings via Metapath2Vec

Results

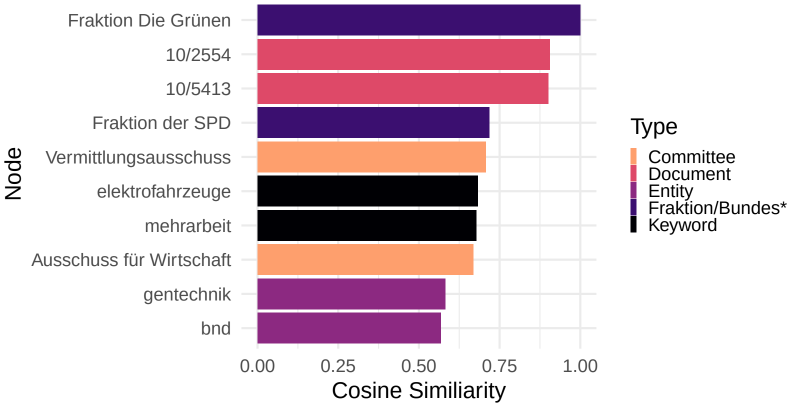


Figure: Visualisation of cosine similarity for "Fraktion Die Grünen". The graphic shows the two most similar nodes with regard to their connection pattern in the network for each node type.

Results

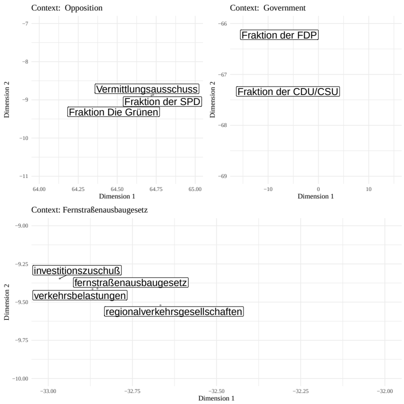


Figure: Visualisation of near neighbours for Opposition / Government and Fernstraßenausbaugesetz. Documents are removed for readability.

Conclusion

- Scalable identification of policy fields in large corpora
- Allows the combination of multiple units of information in one network
- Relatively easy to interprete through co-occurance focus
- Easily sliceable for temporal analysis
- Multiple decision points lead to a dependence on choice of entities, embeddings, clustering algorithms

Thank you for your attention!

Literature I

Aizawa, A. (2003). An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, *39*(1), 45–65. doi: https://doi.org/10.1016/S0306-4573(02)00021-3

Dong, Y., Chawla, N. V., & Swami, A. (2017). metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining* (pp. 135–144). doi: 10.1145/3097983.3098036

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the second international conference on knowledge discovery and data mining* (p. 226–231). AAAI Press.

Falk, N., & Meisinger, N. (2020). Lapps-clarin germaparl metadata enrichment. Retrieved from `https://hdl.handle.net/11022/0000-0007-E969-D`

Fischer, F. (2003). *Reframing public policy: Discursive politics and deliberative practices*. Oxford: Oxford University Press. doi: 10.1093/019924264X.001.0001

Literature II

Freeman, L. C. (2004). *The development of social network analysis: A study in the sociology of science*. Vancouver, BC: Empirical Press. doi: 10.1016/j.socnet.2005.06.004

Grover, A., & Leskovec, J. (2016). node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 855–864). doi: 10.1145/2939672.2939754

Hajer, M. A. (2002). Discourse analysis and the study of policy making. *European Political Science*, *2*(1).

Heclo, H. (1978). Issue networks and the executive establishment. In A. King (Ed.), *The new american political system* (pp. 46–57). Washington: American Enterprise Institute for Public Policy Research.

Heinz, J. P., Laumann, E. O., Salisbury, R. H., & Nelson, R. L. (1990). Inner circles or hollow cores? elite networks in national policy systems. *The Journal of Politics*, *52*(2), 356–390. doi: 10.2307/2131898

Jennings, W., & Bevan, S. (2012, May). *Uk policy agendas codebook.* Retrieved 2021-06-24, from https://policyagendasuk.wordpress.com/codebook/

Literature III

Kenis, P., & Schneider, V. (1991). Policy networks and policy analysis: Scrutinizing a new analytical toolbox. In B. Marin & R. Mayntz (Eds.), *Policy networks* (pp. 25–62). Frankfurt am Main: Campus-Verl.

Laumann, E. O., & Knoke, D. (1987). *The organizational state: Social choice in national policy domains*. Madison, Wis.: The Univ. of Wisconsin Press.

Lazer, D., & Wojcik, S. (2018). Political networks and computational social science. In J. N. Victor, A. H. Montgomery, & M. Lubell (Eds.), *The oxford handbook of political networks* (pp. 115–130). New York, NY: Oxford University Press. doi: 10.1093/oxfordhb/9780190228217.013.9

Leifeld, P. (2018). Discourse network analysis. In J. N. Victor, A. H. Montgomery, & M. Lubell (Eds.), *The oxford handbook of political networks* (pp. 301–325). New York, NY: Oxford University Press. doi: 10.1093/oxfordhb/9780190228217.013.25

Remschel, T., & Kroeber, C. (2020). Every single word: A new data set including all parliamentary materials published in germany. *Government and Opposition*, 1–20. doi: 10.1017/gov.2020.29

Literature IV

Rheault, L., & Cochrane, C. (2020). Word embeddings for the analysis of ideological placement in parliamentary corpora. *Political Analysis*. Retrieved from https://doi.org/10.1017/pan.2019.26 doi: 10.1017/pan.2019.26

Rodriguez, P., & Spirling, A. (2021). Word embeddings: What works, what doesn't, and how to tell the difference for applied research. , *2021*. Retrieved 2021-06-29, from https://www.journals.uchicago.edu/doi/10.1086/715162 (Publisher: The University of Chicago Press) doi: 10.1086/715162

Sabatier, P. A. (1988). An advocacy coalition framework of policy change and the role of policy-oriented learning therin. *Policy Sciences*(23), 129–168.

Sarshar, N., & Roychowdhury, V. (2005). Multiple power-law structures in heterogeneous complex networks. *Physical Review E*, *72*(2), 026114. doi: 10.1103/PhysRevE.72.026114

Stellargraph. (2020). *Comparison of clustering of node embeddings with a traditional community detection method — StellarGraph 1.0.0rc1 documentation.* Retrieved 2021-06-21, from https://stellargraph.readthedocs.io/en/v1.0.0rc1/demos/community_detection/attacks_clustering_analysis.html

Literature V

Sun, Y., & Han, J. (2013). Meta-path-based search and mining in heterogeneous information networks. *Tsinghua Science and Technology*, *18*(4), 329–338. doi: 10.1109/TST.2013.6574671

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, *9*(86), 2579-2605. Retrieved from http://jmlr.org/papers/v9/vandermaaten08a.html

Won, M., & Fernandes, J. M. (2021). Analyzing twitter networks using graph embeddings: an application to the british case. *Journal of Computational Social Science*, 1–11. doi: https://doi.org/10.1007/s42001-021-00128-6

Yanow, D. (2009). *Conducting interpretive policy analysis* ([Nachdr.] ed., Vol. 47). Thousand Oaks, Calif.: Sage Publ. Retrieved from http://www.loc.gov/catdir/enhancements/fy0656/99006588-d.html