Improving Barycentric Embeddings of Topic Spaces

Dora Kiesel[*] Patrick Riehmann[*] Fan Fan[*] Yamen Ajjour[*] Henning Wachsmuth[†]
Benno Stein[*] Bernd Froehlich[*]

[*]Bauhaus-Universität Weimar [†]Paderborn University

Abstract

Topic modeling algorithms such as Latent Dirichlet Allocation (LDA) typically represent documents as a weighted combination of topics. Therefore, generalized barycentric coordinates are a natural fit for the visualization of a topic space. However, spatial positions in a planar barycentric coordinate system are ambiguous for more than three coordinates. Our glyphs for representing documents in combination with layout guidelines help to reduce the positional ambiguity. With an increasing number of documents, barycentric coordinate embeddings suffer from overplotting and visual clutter like other embeddings, possibly even more so since document positions are fully independent of each other. Our experiments with jittering, aggregating glyphs, and grids show potential to reduce these problems for barycentric and other layouts.

Index Terms: Human-centered computing—Visualization—Visualization techniques—Barycentric Coordinates;

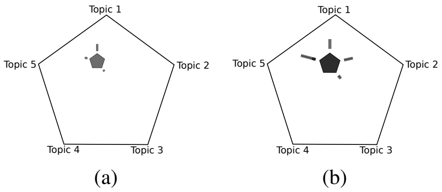
1 Introduction

In order to quickly get an overview of the contents of a text corpus, the visualization of the topics contained in the corpus can be very helpful. Topics can be obtained algorithmically using, for example, Latent Dirichlet Allocation (LDA, Blei [1]): Each document is modeled as a probability distribution over a predefined number of topics, expressing for each topic to what extent the topic is contained within the text of the document.

While such a document model can be visualized by standard means such as scatterplot matrices and parallel coordinates, we propose to use a topic space visualization for topical information. This way, not only the distribution of topics but also the distribution of topic combinations can be assessed visually. Albeit being computationally expensive for large document sets, often multi-dimensional scaling (MDS) or similar embeddings are used to arrange documents according to their topical similarity in a 2D representation of a topic space. However, generalized barycentric coordinates (Meyer, [4]) are a more suitable approach for the document representation generated by an LDA since a document's vector of topic probabilities can directly be used as its barycentric coordinates. The location of the corresponding glyph in the topic space is defined as a weighted linear combination of the vertices of a polygon, which represent the individual topics derived by the LDA (Fig. 1(a)). Unfortunately, positions in the polygonal topic space do not have unique generalized barycentric coordinates – very different topic combinations can result in similar or even the same positions in topic space and, thus, lead to misinterpretations. Similar to other embeddings, overplotting and visual clutter occur more the higher the number of documents is and may be even more pronounced due to the position ambiguity.

2 Related Work

Other works already aim at solving problems that come with generalized barycentric coordinates. Cheng and Mueller [3] experiment with topic space transformations to minimize layout errors. Similarly, Zanabria et al. [7] present a topic space transformation for star coordinates to reduce visual clutter caused by a large number of dimensions. Our work, in contrast, sets the topic space to be a regular polygon and concentrates on enhancing the arrangement and representations of data points. Riehmann et al. [5] introduce glyphs to represent each document that is visualized in a topic space. Each glyph encodes the topic distribution of the respective document (Fig. 1(a)) which reduces position ambiguity but limits scalability to larger document sets due to increased visual clutter and overplotting. Guidelines for designing glyphs can be found in Borgo et al. [2].



(a) (b)

Figure 1: Glyphs design: (a) Using spikes as topic hints as in [5] for a single document. (b) Overplotting of spikes of two documents.

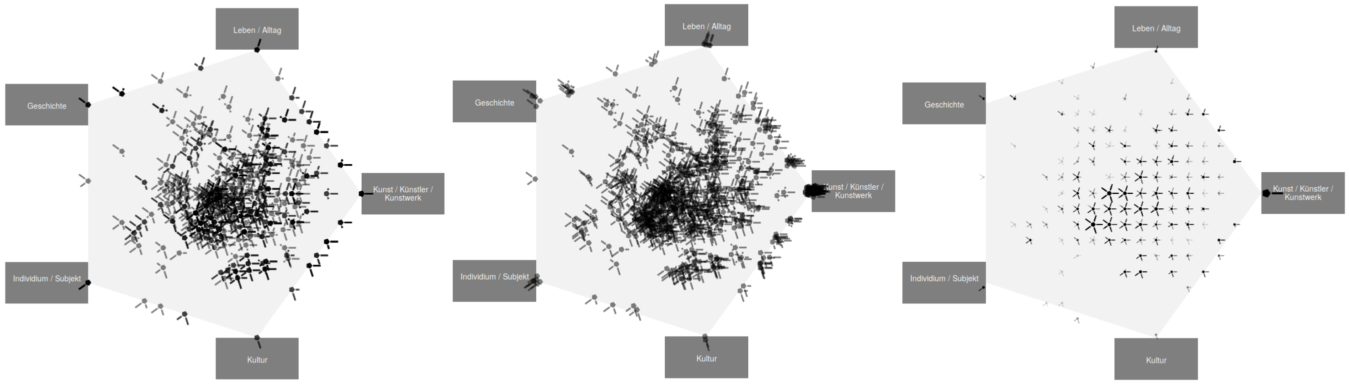
3 Reducing Overplotting, Visual Clutter, and Position Ambiguity

To reduce overplotting in Riehmann's [5] embedding (Fig. 2(a)) and to show how many glyphs are drawn on top of each other, jittering of the glyph positions can be applied (Fig. 2(b)). Jitter could be biased by the direction of the main topic of a document, such that subsets of documents sharing the same main topic are perceived as a cluster.

Alternatively, glyphs placed at the same position and a small neighborhood can be binned using aggregating glyphs, which need to encode the number as well as the topic distribution of the aggregated documents. While the number of documents can easily be encoded by the size of the glyph, the distribution of topics, however, can only be coarsely represented. A simple solution would be to plot all documents on top of each other using alpha blending (Fig. 1(b), Fig. 2(c)). Taking the binning idea further leads to placing a grid over the polygonal topic space and using just a single glyph per grid cell. The resulting tidy display is free of overplotting and visual clutter (Fig. 2(c)). However, the positional ambiguity is increased compared to individual glyphs for each document.

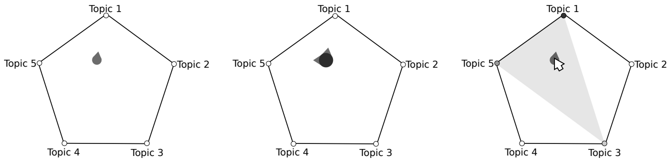
The topic space can be optimized to reduce position ambiguity by choosing the layout and order of the topics appropriately. An even number of topics leads to a hotspot in the center of the polygon since all spatially opposing topics with equal weights map to the center. An uneven number of topics mitigates this problem to a certain extent. For further reduction of these cumulation effects, one should place related topics adjacent to each other. In general,

[*]e-mail: <first>.<last>@uni-weimar.de
[†]e-mail: henningw@upb.de



(a) (b) (c)

Figure 2: Different appearances of the same topics and documents. (a) Regular barycentric visualization. (b) Introducing jitter provides an impression of the actual number of documents. (c) Using a grid and aggregating glyphs leads to a tidier display. Note that glyphs can be positioned slightly outside the gray polygonal area when jitter or a grid layout is used.



(a) (b) (c)

Figure 3: Reduced Design: (a) Single Document. (b) Aggregation of two documents. (c) Hovering reveals a polygon connecting all relevant topics.

the possibility to rearrange topics around the polygonal topic space improves the understanding of topical document clusters.

Additionally, we suggest to use a reduced glyph that encodes only the most prominent topic of the corresponding document (Fig. 3(a)), thus, improving the glyph's legibility. Its distinct appearance also facilitates disambiguating positions. The full topic distribution is visualized on demand by hovering over the document glyph, which reveals a polygon that connects all relevant topics. The reduced design simplifies the readability of aggregating glyphs as well. While the original glyphs utilize alpha blending to retain as many details as possible, the reduced aggregating glyphs implement multiple arrows to point to the most prominent topics contained (Fig. 3(b)), thus, further reducing the position ambiguity.

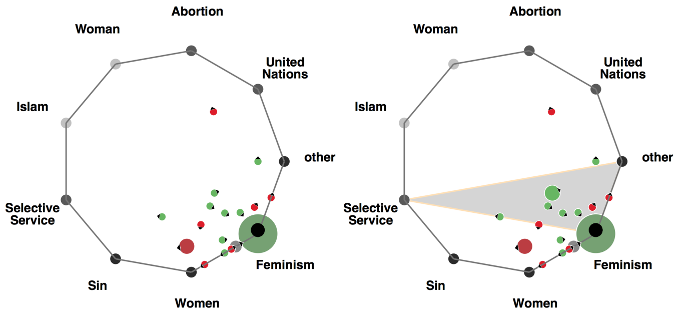
4 Use Case and Future Work

The reduced glyph design is implemented in a visual interface for an argument search engine (https://args.me) [6] (Fig. 4). Each glyph represents one argument of the search result list, is positioned according to its intrinsic topic distribution, and colored according to its stance towards the search query. For the topic space an odd number of topics has been chosen: the eight most prominent ones for the current query plus the ”other” topic that summarizes the rest of the topics of the topic model.

While the system works quite well as it is, we are convinced that using the different visual document representations and animated transitions between them will further support the user in creating a mental model of the topic space and the visualized document set.

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research.*, 3:993–1022, Mar. 2003.



(a) (b)

Figure 4: (a) Interactive visualization of the query results of ”feminism” of an argument search engine (https://args.me), showing the eight most prominent topics of the results plus the ”other” topic — a summarization of all other. The color displays the stance as pro ●, contra ●, and neither ●. The size indicates the number of arguments being represented. (b) Hovering over an argument shows all topics that influenced its position and additionally magnifies the glyph.

[2] R. Borgo, J. Kehrer, D. H. Chung, E. Maguire, R. S. Laramee, H. Hauser, M. Ward, and M. Chen. Glyph-based visualization: Foundations, design guidelines, techniques and applications. In *Eurographics (STARs)*, pp. 39–63, 2013.

[3] S. Cheng and K. Mueller. Improving the fidelity of contextual data layouts using a generalized barycentric coordinates framework. In *Visualization Symposium (PacificVis), 2015 IEEE Pacific*, pp. 295–302. IEEE, 2015.

[4] M. Meyer, A. Barr, H. Lee, and M. Desbrun. Generalized barycentric coordinates on irregular polygons. *Journal of graphics tools*, 7(1):13–22, 2002.

[5] P. Riehmann, D. Kiesel, M. Kohlhaas, and B. Froehlich. Visualizing a thinker's life. *IEEE Transactions on Visualization and Computer Graphics*, 2018. doi: 10.1109/TVCG.2018.2824822

[6] H. Wachsmuth, M. Potthast, K. Al-Khatib, Y. Ajjour, J. Puschmann, J. Qu, J. Dorsch, V. Morari, J. Bevendorff, and B. Stein. Building an Argument Search Engine for the Web. In *Proceedings of the Fourth Workshop on Argument Mining (ArgMining 17)*, 2017. doi: 10.18653/v1/W17-5106

[7] G. G. Zanabria, L. G. Nonato, and E. Gomez-Nieto. istar (i*): An interactive star coordinates approach for high-dimensional data exploration. *Computers & Graphics*, 60:107–118, 2016.