Interactive Curation of Semantic Representations in Digital Libraries

Tim Repke[1(✉)] and Ralf Krestel[2,3]

[1] Hasso Plattner Institute, University of Potsdam, Potsdam, Germany
tim.repke@hpi.uni-potsdam.de
[2] Kiel University, Kiel, Germany
[3] ZBW Leibniz Information Centre for Economics, Kiel, Germany
r.krestel@zbw.eu

Abstract. Digital libraries often contain many heterogeneous documents and cover a variety of topics. Computer generated virtual maps of such collections can help to get an overview and explore the data. The position of each document from the corpus on this virtual two-dimensional map is determined by its semantic similarity to the other documents. However, the computed layout of the data may not adhere to the expectation of domain experts. To this end, we propose a novel approach that enables users to interactively curate the layout of the data. By dragging only a few documents on the canvas, the user can adjust the computed layout to better reflect the expected interpretation of the underlying data. We demonstrate the effectiveness and robustness of our approach using a series of real world datasets.

Keywords: Artificial intelligence for digital libraries · Dimensionality reduction · Data visualisation · Interactive machine learning

1 Introduction

Many digital libraries contain large numbers of heterogeneous documents covering a variety of topics. In order to get an overview and explore these collections, suitable semantic representations are needed to allow intuitive visualisations later. Deep learning methods provide one way to embed documents into semantic spaces, where each document can be represented by a high-dimensional vector. Vectors of semantically similar documents hereby reside closer to one another in this space. In the context of digital libraries, there is a plethora of related work on learning and applying representations for documents. High-dimensional vector representations are utilised in explainable models to interactively gather insights into a digital library [36,37], for visual search interfaces [25,34,38], as well as for interactive clustering or classification [11,12]. Visualisations, such as overview maps of an entire digital library, are powerful tools to explore the data [15]. On such a map of a digital library, which we call the *document landscape*, each document is represented by a point on a two-dimensional canvas,

H.-R. Ke et al. (Eds.): ICADL 2021, LNCS 13133, pp. 219–229, 2021.
https://doi.org/10.1007/978-3-030-91669-5_18

such that semantically similar documents are near one another. The layout of the points on the document landscape is typically done by a dimensionality reduction algorithm that projects the high-dimensional representations into a two-dimensional space, while preserving their pairwise cosine similarities as best as possible. This process has several drawbacks: First, the dimensionality reduction inevitably looses information and usually only favours local similarities. Second, the underlying data may be interpreted in different ways depending on the use-case and fine-tuning the embedding model in an unsupervised setting is also not possible. Once domain experts can explore the layout using an appropriate interactive visualisation, they may be able to suggest edits by dragging documents to different locations to better fit their mental model based on their valuable background knowledge. For example, assuming the documents mostly consist of business reports, a financial expert may want to group documents by industry sectors, whereas an environmental expert may prioritise geographical and technological aspects. The ability to interactively curate the visualisation of embedded documents is easier than, for example, developing a special document embedding model for fiction novels that learns explicitly designed aspects [30]. Therefore, we propose to use an algorithm, that enables users to manipulate the layout with only a few drag-and-drop edits of data-points. The algorithm should update the layout accordingly, while preserving the overall arrangement where possible. This reduces the manual effort to create usable maps of a dataset for a specific use-case, as single edits are augmented. Such an algorithm needs to take the intent behind a user's edit into account. We define an *edit* to be the action of dragging a single point on the map to a new location. There are three fundamental intents, namely *Separate*, *Merge*, and *Arrange*.

In this paper, we propose *ediMAP*, which is based on UMAP [23] to augment the curation of two-dimensional maps of data. The feedback provided by the user by updating the position of a few data points on the map is used to update the underlying similarity graphs and thus the two-dimensional layout of the data. We demonstrate the effectiveness of our approach using several real-world datasets.

2 Related Work

Visualisations are valuable tools to explore digital libraries and gain insights in an intuitive manner. In this paper, we focus on the computer assisted interactive curation of map-like visualisations, which are two-dimensional semantic layouts of a digital library. This form of visualisation has been used on text corpora in different domains, for example in medicine [37], for climate change research [6], and patents [16]. Pang et al. [27] found that transferring concepts and analogies from geographic maps to these artificial maps helps users to get a better overview of their digital library. Depending on the use-case, users may want to be able to interactively manipulate the layout of the data. We identify three ways to incorporate user feedback into the layout process. First, by preconditioning the layout process. Therefore, dimensionality reduction algorithms either use (partially) user annotated data [3,26], manually annotated pairs of very similar or dissimilar items from the digital library [24], or the map is initialised by placing a few items on the empty canvas [31]. Second, by interactive model parametrisation, which allows users to update the layout by changing model parameters [32] or composing a mixture of multiple models [14]. Third, by directly editing existing layouts, where users can manipulate the position of points in an existing two-dimensional layout by dragging.

In this paper, we focus on the third way of incorporating user feedback. Endert et al. [10] discussed interaction patterns for semantic landscapes. In their work, they also proposed a framework of updating a force-based layout of the data. Spathis et al. [33] use a very similar framework. They however use a neural network to first replicate a reference layout provided by an arbitrary dimensionality reduction algorithm. Edits made by a user are then used to update the model. Both these approaches are limited to handle only very small datasets. Contrary to directly editing the landscape, Yuan et al. [18] proposed a dimensionality reduction algorithm that can be influenced by combinations of quality metrics. However, their goal is to optimise visualisations of multi-variate data reduced to more than two dimensions. In either of these setups, a fundamental requirement is the interpretability of the resulting visualisation as discussed by Ding et al. [9]. Furthermore, Lespinats et al. [22] raised the question, whether it is even possible to find faithful two-dimensional mappings of the originally high-dimensional data. Bian et al. [5] avoid this issue by updating the input data itself. Each edit done by a user in the two-dimensional visualisation produced by multi-dimensional scaling of a pre-trained BERT model [8] is propagated back to update the model's last layer.

In our work, we acknowledge the fact, that a layout cannot preserve both, the global and local similarities of all points. Especially in embedding models of textual corpora, there are many ambiguities and overlapping word senses [4] as well as semantic and syntactic subspaces [28]. Others argue that the concept of proximity in a high-dimensional space may not be qualitatively meaningful [1] or that the distribution of distances in this space is closely related, among other characteristics, to the discriminability of the data [17]. Therefore it is essential to identify different interpretations and aspects of the data a user intents to prioritise by their edits. Kobak et al. [19] have demonstrated that dimensionality reduction algorithms can be strongly influenced by their initialisation. User edits could be incorporated in the initialisation to influence the overall result. This approach would be limited to very coarse-grained changes to the overall layout. Most of the recent state-of-the-art algorithms for dimensionality reduction are based on a graph representation of similarity neighbourhoods: either by directly embedding graphs using a neural model [35], manifold approximation [23], or as most recently proposed by minimising distortion functions [2]. We utilise the representation of the data as simplicial sets in UMAP [23] by updating the simplices based on the similarity neighbourhoods affected by user edits.

3 Computer-Assisted Curation of Document Landscapes

The layout of a document landscape of a digital library is typically computed by reducing the dimensionality of high-dimensional semantic representations of all documents from that library. We assume this set of high-dimensional vector representations $x_i \in X \subset \mathbb{R}^n$ to be given. Dimensionality reduction algorithms use a similarity metric, defined by a distance measure $d(x_i, x_j)$ between high-dimensional points, to compute a projection that faithfully preserves pairwise similarities in the layout. This projection $P : \mathbb{R}^n \mapsto \mathbb{R}^2$ maps each item x_i to its two-dimensional counterpart $y_i \in Y$. For the scope of this work, we assume that such an *initial layout* already exists. Given the *initial layout* of the digital library, a user curates the map by dragging points from their *source location* y_k to their *target location* $\mathring{y}_k$. The proposed algorithm assists this curation process by using these changes and computing an *updated layout* of the digital library. Thereby, the updated layout should fulfil the following set of objectives.

1. A manipulated point should be positioned at or near the target location in the updated layout.
2. The position of points in the proximity of the target location shall not differ significantly from the initial layout.
3. Points in the proximity of the source location may also move to a different location if necessary.
4. Changes to the general layout of the map shall be minimal to preserve the user's mental map of the data. This can also be referred to as stability or robustness.

Note, that algorithms for assisting the curation process may not be able to accommodate all these objectives, as the actually intended edit may contradict some of these objectives.

Our proposed algorithm assisting the curation of document landscapes is based on the popular UMAP algorithm [23]. The core principle of UMAP is to use a network of similar items from the dataset to create the two-dimensional layout of the data. The similarities are represented as a weighted network of items that are neighbours in the high-dimensional space, so-called fuzzy simplicial sets. A spectral embedding of this network is used to initialise the layout, which is then fine-tuned with a force-directed layout algorithm.

We utilise this concept of manifold approximation and force-directed graph drawing in *ediMAP*. Hereby, the edits to the document landscape suggested by a user are used to update the similarity network. As shown by related work, the starting point of any (re-)layout process has a significant impact on the resulting layout [19,29]. By continuing to update the initial layout and only partially changing the underlying similarities, we preserve the mental map of the data as much as possible. Since we assume that only the layout of a digital library is provided and not a fitted UMAP model, we first need to construct the normalised similarity graph. The similarities are based on the distance measure $d(x_i, x_j)$ between high-dimensional vectors, each representing a respective document from the library. Let $\mathcal{N}^k_{x_i}$ be the set of k nearest neighbours of document x_i. For each $x_j \in \mathcal{N}^k_{x_i}$, we add an edge (x_i, x_j) to the similarity graph. The edge weights are defined as $\exp(-\max\{0, d(x_i, x_j) - \rho\}/\sigma)$, where ρ is the distance to the closest neighbour to x_i and σ the distance to the k-th closest neighbour to x_i. Note, that UMAP is actually defining σ to be the smoothed k nearest neighbour distance.

This similarity graph is based on the high-dimensional vector representations, but should also reflect the proximity between documents in the initial layout, as most dimensionality reduction algorithms commonly aim to preserve local similarity neighbourhoods.

As a user moves a document x_i from its source location y_i in the initial layout to its target location $\mathring{y}_i$, we update the similarity graph as follows. First, we determine the k nearest neighbours $\mathcal{N}^k_{\mathring{y}_i}$ of the manipulated document at the target location $\mathring{y}_i$. As before, we add edges for each neighbour to the similarity graph and weigh them by their normalised distance. This time, however, we use the average of the normalised distances in the high-dimensional and two-dimensional space. Using only either one space to determine the similarity weight would either neglect the actual similarity in the semantic representation or what the user actually sees while curating the document landscape. Furthermore, we update the edge weights of the original neighbourhood of the manipulated document x_i in the similarity graph as follows. Edges, if they exist, connecting x_i to the k nearest neighbours $x_j \in \mathcal{N}^k_{y_i}$ at the source location y_i, are reduced by the factor $\xi \in (0, 1)$. All edges, apart from the aforementioned, connecting any $x_j \in \mathcal{N}^k_{y_i}$ are reduced by the factor of ξ^2. This reduction of edge weights limits the otherwise counteracting forces in the update phase of the layout. Furthermore, it could be exposed in a user interface for curating document landscapes as a user defined parameter to influence, how much the documents in the source neighbourhood should be moved along with the document that was edited.

Finally, we revise the layout of the document landscape using a force-directed layout algorithm based on the updated similarity graph. For each node that was affected by updating the similarity graph, we iteratively update the location of the respective document's position y_i on the landscape over several epochs. In each epoch of the layout optimisation, the location is updated to $\tilde{y}_i$ using

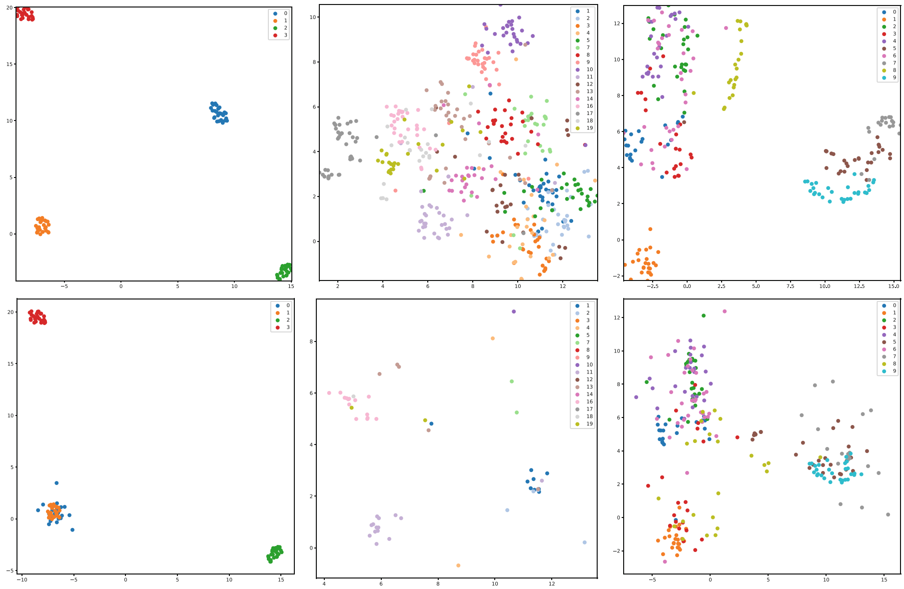
$$\tilde{y}_i = y_i - \eta \sum_{(y_i, y_j, w_{i,j}) \in \mathcal{G}} w_{i,j} \frac{(y_i - y_j)}{\|y_i - y_j\|},$$

where η is the learning rate, which decays with each iteration. Typically, force directed layout algorithms require an additional repelling force. However, since we only manipulate the locations of documents affected by the update of the similarity graph and use all the remaining that are connected to these documents in the graph as fixed references, we only need the attracting force defined above.

4 Evaluation

In this section, we apply our model for interactive dimensionality reduction to several real world datasets. We simulate user interactions to measure how well our model can fulfil the expectations and objectives we defined earlier across several different setups. The resulting maps of the datasets are quantitatively and qualitatively evaluated in a series of experiments.

For our experiments, we use six datasets with different characteristics: real world datasets with multivariate, image, and text data, as well as an artificial dataset. This includes the well-known *MNIST*[1] dataset of written digits [21] and the *MNIST-1D*[2] variant, which is derived from the original data but harder to separate [13]. We also use *FashionMNIST*[3], which contains greyscale images of fashion articles like shoes and sweaters across ten categories [39]. Aside from image data, we also use the multivariate *Seeds*[4] dataset [7]. It is comprised of measurements of wheat kernels and thus provides intuitively interpretable dimensions. Furthermore, we evaluate our approach on textual data using the *20-Newsgroups*[5] dataset [20]. Real-world datasets often contain overlapping or ambiguous latent aspects, which makes them difficult to use for evaluation. Thus we generate the artificial *Blobs* dataset to control the latent aspects within the high-dimensional space.



(a) Blobs (b) 20-Newsgroups (tf-idf) (c) FashionMNIST

Fig. 1. Scatterplots of dataset layouts before (top) and after (bottom) curation.

We simulate user edits by moving one or more points with a specific label towards the centroid of points with another label. In this way, we mimic the user intent of merging two clusters of points. This procedure is repeated for different numbers of manipulated points and different sets of labels. In particular,

[1] http://yann.lecun.com/exdb/mnist/.
[2] https://github.com/greydanus/mnist1d.
[3] https://github.com/zalandoresearch/fashion-mnist.
[4] https://archive.ics.uci.edu/ml/datasets/seeds.
[5] http://qwone.com/~jason/20Newsgroups/.

we simulate the following sets of edits: For the *20-Newsgroups* dataset, we simulate edits with the intent to create a document landscape of four clusters of messages. These four clusters are based on the respective subtopics of *computer* (`comp.*`), *recreational* (`rec.*`), *science* (`sci.*`), and *talk* (`talk.*`). Similar to the *20-Newsgroups* dataset, we simulate a curation of merging clusters for the *FashionMNIST* dataset. Of the originally more fine-grained labels, the intent is to form groups of articles, in particular *footwear* (sneaker, boot, sandal), tops (tshirt, pullover, shirt, coat), and others (trouser, dress, bag). Lastly, for the *Blobs* dataset, we simulate the intended merging of two clusters.

Note, that in an actual interface for curating document landscape, the data may not be annotated as described here. We only use these datasets for a clear definition of semantically similar groups of items in a dataset. Figure 1 shows the three datasets before and after applying edits and adapting the layout with *ediMAP*. The artificially generated *Blobs* dataset has a clear separation between differently labelled points in the initial layout. *ediMAP* is able to perfectly achieve the goal of merging the orange and blue clusters without altering the location of any other points. When moving many points from different source locations, as done for the *20-Newsgroups* and *FashionMNIST* datasets, the individual intents are hard to distinguish and many points overlap in the resulting layout. To circumvent this issue, we split the edits across smaller batches and do repeated partial updates. In these settings, adding a repulsion factor as in traditional force directed layout algorithms may resolve this issue.

We also evaluate our approach quantitatively and compare the results to *iSP*, a neural network based approach that learns to replicate the initial layout and can be retrained after edits [33]. We improved the performance by adding another term to the objective function of *iSP* in order to stabilise the layouts.

Table 1. Displacement measures after updating the layout with *iSP* (left) and *ediMAP* (right) based on simulated edits of 10% of points.

	Blobs		News		Seeds		MNIST		MNIST1D		F-MNIST	
Total	0.41	0.08	0.50	0.01	0.45	0.13	0.43	0.06	0.57	0.06	0.36	0.07
Target	0.39	0.09	0.55	0.00	0.51	0.25	0.45	0.08	0.59	0.08	0.33	0.10
Dist	0.43	0.39	0.60	0.26	0.58	0.22	0.49	0.17	0.62	0.13	0.39	0.24
DTT	–0.39	–0.03	–0.27	–0.06	–0.23	–0.04	–0.29	–0.01	–0.12	–0.02	–0.31	–0.02

Corresponding to the previously defined objectives for the updated layout, we determine (1) the distance of the manipulated points between the target location and their location in the updated layout (Dist); (2) the displacement of points between the initial layout and the updated layout of points near the target location (Target); (3) the difference between all pairwise distances of all points in the source and target cluster, based on the label information, before and after updating the layout (DTT); and (4) the total displacement of all points between the two layouts (Total). Smaller numbers are better, DTT should be negative in most cases. However, since the overall purpose of our approach is to update the layout, some movement has to necessarily occur. Thus, we exclude all points from the calculations that were explicitly edited. All results are averaged over multiple runs and normalised by the relevant total number of points and the size of the respective landscape. We move randomly selected points that have the same label towards the centroid of points that have a different label and repeat this process for multiple pairs of labels. All values are normalised to reduce the effect of different sizes of the initial layouts and the varying sizes of the datasets. We list the results for all metrics and configurations after updating the layout in Table 1. Here we can see, that the layouts updated using *ediMAP* generally cause less displacement to the overall layout, as shown by the TOTAL and TARGET metrics. Although *iSP* uses a mask to minimise the effect it has on the general layout, there is still more movement overall. Both algorithms show a similar performance in the DTT and DIST metrics, however *ediMAP* requires less points to be edited to minise these numbers.

In conclusion, we were able to demonstrate that our proposed *ediMAP* algorithm provides useful assistance for curating a document landscape. In judging the performance of *ediMAP* or any other curation assistance it is important to note, that identifying the intent of a user's edit is almost impossible. Using only the feedback of dragging a document to a new target location can be interpreted in many different ways. With *ediMAP*, we focused on one aspect, the intent of merging clusters of documents a user determined to be similar and were able to show its effectiveness to achieve that goal.

5 Conclusion

In this paper, we presented an approach for the interactive curation of semantic representations in digital libraries. Given a two-dimensional projection, which we call the document landscape, our algorithm is able to assist the curation process based on only a few suggested edits by a user. We described, how our *ediMAP* algorithm uses a similarity graph to update existing layouts of document landscapes. Additionally, we improved a neural network based model from related work to use as a baseline. In the evaluation on several real-world datasets, we were able to demonstrate the effectiveness of our approach.

However, as discussed in the evaluation, identifying the user's edit intent is important, yet very challenging. Conditioning models on specific intents to generate several suggested updated document landscapes for a given edit will be part of future work, along with an actual user interface. Intuitive and semantically meaningful visualisations of digital libraries heavily rely on good high-dimensional semantic representations of the documents. Utilising the user feedback given during the curation process could be propagated back to a representation model to improve the model itself, not just the 2-dimensional visualisation.

References

1. Aggarwal, C.C., Hinneburg, A., Keim, D.A.: On the surprising behavior of distance metrics in high dimensional space. In: Van den Bussche, J., Vianu, V. (eds.) ICDT 2001. LNCS, vol. 1973, pp. 420–434. Springer, Heidelberg (2001). https://doi.org/10.1007/3-540-44503-X_27
2. Agrawal, A., Ali, A., Boyd, S.P.: Minimum-distortion embedding (2021). arXiv:2103.02559
3. An, S., Hong, S., Sun, J.: Viva: semi-supervised visualization via variational autoencoders. In: Plant, C., Wang, H., Cuzzocrea, A., Zaniolo, C., Wu, X. (eds.) Proceedings of the International Conference on Data Mining (ICDM), pp. 22–31. IEEE (2020). https://doi.org/10.1109/ICDM50108.2020.00011
4. Arora, S., Li, Y., Liang, Y., Ma, T., Risteski, A.: Linear algebraic structure of word senses, with applications to polysemy. Trans. Assoc. Comput. Linguist. (TACL) **6**, 483–495 (2018)
5. Bian, Y., North, C.: DeepSI: interactive deep learning for semantic interaction. In: Proceedings of the International Conference on Intelligent User Interfaces (IUI), pp. 197–207. ACM Press, Geneva, Switzerland (2021)
6. Callaghan, M., Minx, J., Forster, P.: A topography of climate change research. Nat. Clim. Change **10**, 118–123 (2020). https://doi.org/10.1038/s41558-019-0684-5
7. Charytanowicz, M., Niewczas, J., Kulczycki, P., Kowalski, P.A., Łukasik, S., Żak, S.: Complete gradient clustering algorithm for features analysis of x-ray images. In: Piętka, E., Kawa, J. (eds.) Information Technologies in Biomedicine. Advances in Intelligent and Soft Computing, vol. 69, pp. 15–24. Springer, Berlin, Heidelberg (2010). https://doi.org/10.1007/978-3-642-13105-9_2
8. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), pp. 4171–4186. ACL (2019). https://doi.org/10.18653/v1/n19-1423
9. Ding, J., Condon, A., Shah, S.P.: Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. Nat. Commun. **9**(1), 1–13 (2018)
10. Endert, A., Fiaux, P., North, C.: Semantic interaction for visual text analytics. In: Proceedings of the SIGCHI conference on Human Factors in Computing Systems (CHI), pp. 473–482. ACM Press (2012)
11. Ghosal, T., Raj, A., Ekbal, A., Saha, S., Bhattacharyya, P.: A deep multimodal investigation to determine the appropriateness of scholarly submissions. In: Proceedings of the Joint Conference on Digital Libraries (JCDL), pp. 227–236. IEEE (2019). https://doi.org/10.1109/JCDL.2019.00039
12. Ghosal, T., Sonam, R., Ekbal, A., Saha, S., Bhattacharyya, P.: Is the paper within scope? Are you fishing in the right pond? In: Proceedings of the Joint Conference on Digital Libraries (JCDL), pp. 237–240. IEEE (2019). https://doi.org/10.1109/JCDL.2019.00040
13. Greydanus, S.: Scaling *down* deep learning. CoRR abs/2011.14439 (2020). arXiv:2011.14439
14. Hilasaca, G.M.H., Paulovich, F.V.: User-guided dimensionality reduction ensembles. In: Proceedings of the International Conference on Information Visualisation (IV), pp. 228–233. IEEE (2019)

15. Hogräfer, M., Heitzler, M., Schulz, H.J.: The state of the art in map-like visualization. In: Computer Graphics Forum, vol. 39, pp. 647–674. Wiley Online Library (2020)
16. Hoo, C.S.: Impacts of patent information on clustering in derwent innovation's themescape map. World Pat. Inf. **63**, 102001 (2020). https://doi.org/10.1016/j.wpi.2020.102001
17. Houle, M.E.: Dimensionality, discriminability, density and distance distributions. In: Proceedings of the International Conference on Data Mining (ICDM), pp. 468–473. IEEE (2013). https://doi.org/10.1109/ICDMW.2013.139
18. Johansson, S., Johansson, J.: Interactive dimensionality reduction through user-defined combinations of quality metrics. Trans. Vis. Comput. Graph. (TVCG) **15**(6), 993–1000 (2009)
19. Kobak, D., Linderman, G.C.: Initialization is critical for preserving global data structure in both t-SNE and UMAP. Nat. Biotechnol. 1–2 (2021)
20. Lang, K.: Newsweeder: learning to filter netnews. In: Proceedings of the International Conference on Machine Learning (ICML), pp. 331–339. Morgan Kaufmann (1995). https://doi.org/10.1016/b978-1-55860-377-6.50048-7
21. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
22. Lespinats, S., Aupetit, M.: CheckViz: sanity check and topological clues for linear and non-linear mappings. Comput. Graph. Forum **30**(1), 113–125 (2011)
23. McInnes, L., Healy, J., Saul, N., Großberger, L.: UMAP: uniform manifold approximation and projection. J. Open Source Softw. **3**(29), 861 (2018)
24. Meng, M., Wei, J., Wang, J., Ma, Q., Wang, X.: Adaptive semi-supervised dimensionality reduction based on pairwise constraints weighting and graph optimizing. Int. J. Mach. Learn. Cybern. **8**(3), 793–805 (2015). https://doi.org/10.1007/s13042-015-0380-3
25. Mesbah, S., Fragkeskos, K., Lofi, C., Bozzon, A., Houben, G.-J.: Facet embeddings for explorative analytics in digital libraries. In: Kamps, J., Tsakonas, G., Manolopoulos, Y., Iliadis, L., Karydis, I. (eds.) TPDL 2017. LNCS, vol. 10450, pp. 86–99. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67008-9_8
26. Mikalsen, K.Ø., Soguero-Ruíz, C., Bianchi, F.M., Jenssen, R.: Noisy multi-label semi-supervised dimensionality reduction. Pattern Recogn. **90**, 257–270 (2019). https://doi.org/10.1016/j.patcog.2019.01.033
27. Pang, P.C.I., Biuk-Aghai, R.P., Yang, M., Pang, B.: Creating realistic map-like visualisations: results from user studies. J. Vis. Lang. Comput. (JVLC) **43**, 60–70 (2017)
28. Reif, E., et al.: Visualizing and measuring the geometry of BERT. In: Proceedings of the Conference on Neural Information Processing Systems (NIPS), pp. 8592–8600. NIPS Foundation Inc., San Diego, USA (2019). https://proceedings.neurips.cc/paper/2019/hash/159c1ffe5b61b41b3c4d8f4c2150f6c4-Abstract.html
29. Repke, T., Krestel, R.: Robust visualisation of dynamic text collections: measuring and comparing dimensionality reduction algorithms. In: Proceedings of the Conference for Human Information Interaction and Retrieval (CHIIR) (2020)
30. Risch, J., Garda, S., Krestel, R.: Book recommendation beyond the usual suspects. In: Dobreva, M., Hinze, A., Žumer, M. (eds.) ICADL 2018. LNCS, vol. 11279, pp. 227–239. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-04257-8_24
31. Saket, B., Endert, A., Rhyne, T.: Demonstrational interaction for data visualization. IEEE Comput. Graph. Appl. **39**(3), 67–72 (2019)

32. Shi, X., Yu, P.S.: Dimensionality reduction on heterogeneous feature space. In: Proceedings of the International Conference on Data Mining (ICDM), pp. 635–644. IEEE (2012). https://doi.org/10.1109/ICDM.2012.30
33. Spathis, D., Passalis, N., Tefas, A.: Interactive dimensionality reduction using similarity projections. Knowl.-Based Syst. **165**, 77–91 (2019)
34. Vahdati, S., Fathalla, S., Auer, S., Lange, C., Vidal, M.-E.: Semantic representation of scientific publications. In: Doucet, A., Isaac, A., Golub, K., Aalberg, T., Jatowt, A. (eds.) TPDL 2019. LNCS, vol. 11799, pp. 375–379. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30760-8_37
35. Wang, Y., Jin, Z., Wang, Q., Cui, W., Ma, T., Qu, H.: DeepDrawing: a deep learning approach to graph drawing. Trans. Vis. Comput. Graph. (TVCG) **26**(1), 676–686 (2020)
36. Wawrzinek, J., Balke, W.-T.: Measuring the semantic world – how to map meaning to high-dimensional entity clusters in PubMed? In: Dobreva, M., Hinze, A., Žumer, M. (eds.) ICADL 2018. LNCS, vol. 11279, pp. 15–27. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-04257-8_2
37. Wawrzinek, J., Hussaini, S.A.R., Wiehr, O., Pinto, J.M.G., Balke, W.: Explainable word-embeddings for medical digital libraries - a context-aware approach. In: Proceedings of the Joint Conference on Digital Libraries (JCDL), pp. 299–308. ACM Press (2020). https://doi.org/10.1145/3383583.3398522
38. Wohlmuth, C., Correia, N.: User interface for interactive scientific publications: a design case study. In: Doucet, A., Isaac, A., Golub, K., Aalberg, T., Jatowt, A. (eds.) TPDL 2019. LNCS, vol. 11799, pp. 215–223. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30760-8_19
39. Xiao, H., Rasul, K., Vollgraf, R.: Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. CoRR abs/1708.07747 (2017). arXiv:1708.07747