Vision, Modeling, and Visualization (2021)
B. Andres, M. Campen, and M. Sedlmair (Eds.)

Visualizing Temporal-Thematic Patterns in Text Collections

M. Knabben[1] and M. Baumann[1] and T. Blascheck[1] and T. Ertl[1] and S. Koch[1]

[1]University of Stuttgart, Institute for Visualisation and Interactive Systems, Germany

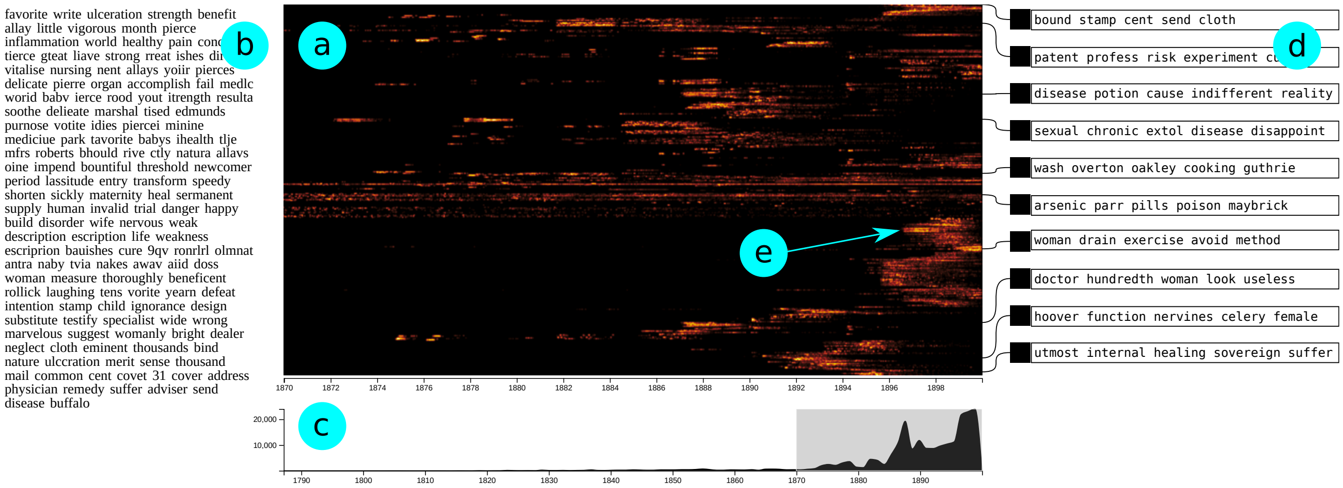

favorite write ulceration strength benefit
allay little vigorous month pierce
inflammation world healthy pain cond
tierce gteat liave strong rreat ishes di
vitalise nursing nent allays yoiir pierces
delicate pierre organ accomplish fail medlc
worid babv ierce rood yout itrength resulta
soothe delieate marshal tised edmunds
purnose votite idies piercei minine
mediciue park tavorite babys ihealth tlje
mfrs roberts bhould rive ctly natura allavs
oine impend bountiful threshold newcomer
period lassitude entry transform speedy
shorten sickly maternity heal sermanent
supply human invalid trial danger happy
build disorder wife nervous weak
description escription life weakness
escriprion bauishes cure 9qv ronrlrl olmnat
antra naby tvia nakes awav aiid doss
woman measure thoroughly beneficent
rollick laughing tens vorite yearn defeat
intention stamp child ignorance design
substitute testify specialist wide wrong
marvelous suggest womanly bright dealer
neglect cloth eminent thousands bind
nature ulccration merit sense thousand
mail common cent covet 31 cover address
physician remedy suffer adviser send
disease buffalo
b
a
d
e
bound stamp cent send cloth
patent profess risk experiment cu
disease potion cause indifferent reality
sexual chronic extol disease disappoint
wash overton oakley cooking guthrie
arsenic parr pills poison maybrick
woman drain exercise avoid method
doctor hundredth woman look useless
hoover function nervines celery female
utmost internal healing sovereign suffer
1870
1872
1874
1876
1878
1880
1882
1884
1886
1888
1890
1892
1894
1896
1898
c
20,000
10,000
1790
1800
1810
1820
1830
1840
1850
1860
1870
1880
1890

Figure 1: *An overview of temporal-thematic patterns of a text collection* a *using our developed prototype. This example collection consists of 217,141 texts about Dr. Pierce's Favorite Prescription. The horizontal position relates to the time stamp, and the vertical position is a 1D embedding of each text. This creates regions with a varying density of texts, which we depict using a color scale based on the black-body radiation (black: no texts, white: most texts). On the left* b*, we list significant terms for one selected location* e *in the diagram and on the right* d *for a sample of regions across the vertical. A line chart* c *shows the number of texts over time and the temporal region that is currently selected (gray rectangle).*

Abstract

Visualizing the temporal evolution of texts is relevant for many domains that seek to gain insight from text repositories. However, existing visualization methods for text collections do not show fine-grained temporal-thematic patterns. Therefore, we developed and analyzed a new visualization method that aims at uncovering such patterns. Specifically, we project texts to one dimension, which allows positioning texts in a 2D diagram of projection space and time. For projection, we employed two manifold learning algorithms: the self-organizing map (SOM) and UMAP. To assess the utility of our method, we experimented with real-world datasets and discuss the resulting visualizations. We find our method facilitates relating patterns and extracting associated texts beyond what is possible with previous techniques. We also conducted interviews with historians to show that our prototypical system supports domain experts in their analysis tasks.

CCS Concepts

• ***Applied computing*** → *Document searching;* • ***Information systems*** → *Search interfaces;* • ***Human-centered computing*** → ***Visualization techniques;***

1. Introduction

This work tackles the challenge of visualizing temporal-thematic patterns in text collections. With the term pattern, we refer to temporal or thematic change in a text collection and in the visualization. We seek to transform data characteristics into visible patterns that show, for example, an increase, a decrease, a constant, or a repeated occurrence of themes in a collection. The scatter plot is a natural way to represent two-dimensional (2D) projected

DOI: 10.2312/vmv.20211366

data. Then, the viewer can see 2D clusters of points their relations. However, scatter plots do not represent temporal information inherently. Therefore, a 2D scatter plot of projected texts is not suitable for our objective. Our work adapts this approach to one-dimensional (1D) projections with time as the second axis. Stream graphs [HHWN02] are another solution for this challenge. Stream graphs depict the number of items (e.g., words or topics) over time by mapping the number to the height of the stream and stacking several of them. However, we find that stream graphs lead to visual clutter if the number of streams overly increases and the individual timelines become barely visible. Therefore, there is a need to develop new methods to inspect temporal-thematic patterns in text collections. Our solution is to project the data to one dimension, use the other dimension for time, and further scale down elements using a pixel-based diagram. Figure 1(a) shows an example of our solution. By placing each document on a vertical position using the projected value and on a horizontal position using time, we enable the viewer to interpret texts according to content and time. Thus, we build on the established principle of showing projected data in scatter plots using a temporal variant.

This work makes the following contributions: We present a novel method and design for visualizing temporal-thematic patterns in text collections. Our diagram is a combination of existing techniques and can show unforeseen patterns that emerge from a collection of texts. We assess the diagram by experimenting with collections containing texts of different types and sizes. The objective of the experiments is to find out whether there are meaningful patterns and limits of the method regarding the size and type of the texts. To assess the utility of this method further, we interview two historians to find out whether they can read the diagram and whether our prototype supports them in their research.

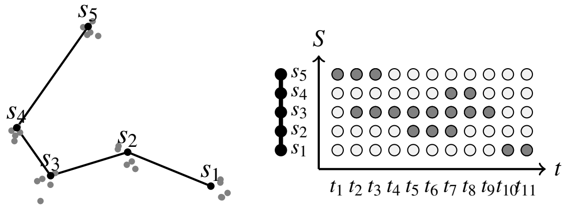
2. Related Work

The method we propose employs a combination of unsupervised learning and visualization. Therefore, we review works that apply clustering and projection techniques and methods for temporal (text) visualization.

2.1. Processing and Dimensionality Reduction

The first step of processing textual data for unsupervised learning is to convert texts to vectors. Two common approaches for vector representations are the vector space model (VSM) [SWY75] and embedding methods (e.g., word2vec or doc2vec). We chose the VSM because it has successfully been used for document clustering and allows an intuitive interpretation of the vectors.

One candidate for projection is uniform manifold approximation and projection for dimension reduction (UMAP) because it allows non-linear projections that do not require the initialization of a distance matrix. Other methods (e.g., PCA [Jol11] or NMF [XLG03]) project data linearly. However, a linear model lacks the precision we need for textual data. This argument holds for topic modeling algorithms such as LDA [BNJ03] in the probabilistic setting as well, as there would be just one weight for representing textual variability. Other non-linear projection methods (e.g., multidimensional scaling (MDS) [CC00] or t-SNE [vdMH08]) are not feasible for large collections (i.e., more than 10,000 elements), because they require the initialization of a pairwise distance matrix.


s_5
s_4
s_3
s_2
s_1
S
s_5
s_4
s_3
s_2
s_1
t
t_1 t_2 t_3 t_4 t_5 t_6 t_7 t_8 t_9 t_{10} t_{11}

(a) *Fitting of linear SOM.* **(b)** *Cartesian grid.*

Figure 2: *First, we represent texts (exemplified as gray dots in Figure a) in a feature space. Then, we fit a linear SOM (connected cells in black) to the text representations. To generate the diagram, we place each text in a 2D diagram in which the vertical axis denotes the cells $S = (s_1, \ldots, s_5)$ and the horizontal axis denotes the date of texts (Figure b).*

Vector quantization seem worthwhile for processing large datasets because they allow batch processing. However, k-means clustering is not a solution because it requires a strict assignment of elements to clusters. The self-organizing map (SOM) [Koh00] is a method suitable for clustering that overcomes this problem. RZESZUTEK, ANDROUTSOS, and KYAN [RAK10] consider the SOM for representing textual data in slices for each time step. SCHRECK, BERNARD, LANDESBERGER, and KOHLHAMMER [SBLK09] use the SOM to incorporate time for the analysis of trajectory data. SARLIN [Sar13] proposes a variant of the SOM that includes time by training a SOM for each time step and connecting these using short-term memory. Applied in our setting, this creates ambiguous patterns as a shift in the vertical location may be due to relocations of cells or a change of data.

2.2. Visualization of Time-oriented Data

Researchers propose time-oriented visualization approaches for depicting the change of data values over time. Some methods display individual time series as geometric shapes, in which the value at a point in time becomes the height of the shape [HHWN02; LYK*12; VWD04]. These methods have in common that individual timelines take up varying vertical space. If one wants to increase the scalability by showing more timelines, one has to aggregate them. A solution for this problem is the hierarchical stream graph [DYW*13]. Using interaction, the viewer can inspect the text collection on different levels of detail. In contrast, our method tries to show as many details as possible at once, with interaction happening later in the analysis.

We know from empirical research that position is a powerful visual variable [CM84]. Hence, we discuss methods that use position to convey (semantic) similarity. BACH, SHI, HEULOT, et al. [BSH*16] represented sequential data by projecting the elements of the sequence in a 2D diagram using MDS. This work is related, however, not applicable to our type of data because we aim at showing the thematic change of text collections. Although texts have a temporal component, there is no inherent sequence of texts.

2.3. Text Visualization

Some methods visualize text collections by placing words in a diagram. Parallel Tag Clouds [CVW09] display sets of words in vertical arrangements. Words are linked across sets which allows comparing word usage in the sets. PyramidTags [KKE20] improves on the tag cloud idea by considering the relatedness of tags, word order, and time. In contrast, the Word Tree [WV08] makes the relations explicit by connecting consecutive words from a text. An advantage of these methods is that words immediately inform the viewer about the content of the texts. However, these approaches have limits because readable words take up more screen space than geometric shapes or even smaller visual elements like pixels. The latter approaches, in contrast, need interaction to inform about the texts.

Researchers investigate the combination of topic modeling and interactive visualization intensely. Some works integrate humans in the modeling loop [RSB*17; ESD*18] to interactively steer the learning process. Other works use topic modeling to represent texts and their temporal evolution [CLT*11]. Their work uses a visualization technique comparable to a stream graph in which a single text influences the height of multiple streams. In contrast, the method presented here represents texts as a single value mapped to a vertical position. A priori, such a difference does not entail different analyses. However, the combination of stream graphs and topic modeling creates an additional abstraction: The viewer cannot distinguish whether changes in multiple streams are due to topic changes within the same text or from different texts. In contrast, projecting text to a single position communicates to the viewer that there are texts with specific content.

2.4. Pixel-based Visualization

Pixel-based methods have the advantage that one can display more information than with rendered geometric shapes [JS98; Kei00]. It is even possible to have a higher pixel density than the screen resolution, as these methods arrange rows or columns such that patterns across individual pixels become visible [Kei00]. A central related work regarding pixel-based methods is MotionRugs. BUCHMÜLLER, JÄCKLE, CAKMAK, et al. [BJC*19] represent spatio-temporal data by mapping the spatial component to one dimension using space-filling curves. Then, they represent the temporal component in the horizontal direction. Structurally, MotionRugs is a similar visualization as the viewer can analyze temporal structures of multivariate data. However, a space-filling curve visits positions in the data space evenly, whereas a projection method tries to capture only regions where data is available. As the likelihood of a single word occurring in a real-world text depends on other words appearing in the text, it is more likely that textual data looks like Figure 2a. Thus a projection method as described in this paper seems more appropriate to our problem.

JÄCKLE, FISCHER, SCHRECK, and KEIM [JFSK16] developed a visual analytics (VA) method for temporal multivariate data using a pixel-based overview. To create the temporal overview, they combine a sliding window method with 1D MDS. This method creates ambiguous patterns in the same way as in the method by SARLIN [Sar13]: A shift in the location can either be due to a change of the data over time or the result of different projections for the time step. As we want to visualize how texts progress over time, we need to eliminate other influences on the vertical position.

In our approach, we project texts using SOM or UMAP to one dimension, combine a text's position along this projection dimension with its time stamp, and depict the density of texts in a pixel-based Cartesian diagram as a color-coded frequency value. This shows the temporal-thematic patterns of text collections. The related works considered in this section tackle similar problems or employ similar methods. However, our approach is a novel combination of these methods that allows seeing fine-grained textual change and global change simultaneously.

3. Method

In this section, we describe the processing steps: text processing, text representation, machine learning, and visual mapping. UMAP allows us to directly project texts to 1D because this algorithm is designed deliberately for such a task. As an alternative, the SOM is fundamentally a clustering algorithm because it assigns data to discrete cells. Therefore, we need a parameterization on the linear structure to obtain a 1D value.

3.1. Processing and Representation of Texts

First, we introduce our used variables for this work: n is the size of the text collection; m is the size of the text collection for training; q is the size of the vocabulary; $f_{\min}$ and $f_{\max}$ are the minimum and maximum document frequencies; k is the number of SOM cells; T is the number of training steps; r is the initial radius of the SOM; h and w are the height and width of the diagram in pixels.

In the first step, we abstract from textual data. We assume there is a collection of texts including a temporal component $C' = \{(d_i,t_i)\}_{i=1}^{n}$, where d_i is a text in a natural language, and t_i is the time (e.g., the publication date of the text). We process the texts in a way that is common in natural language processing (NLP): tokenize the strings, then pos-tag each token, and finally, lemmatize verbs, nouns, adjectives, and adverbs. Then, we create feature vectors, for which the value of each feature is the relative term frequency weighted by the inverse document frequency (tf-idf) [WLWK08]. To reduce the computing time, we remove infrequent terms in the texts (i.e., the document frequency for a term is lower than $f_{\min}$). Terms that appear in (almost) all texts (i.e., the share of texts that contains a term is higher than $f_{\max}$) do not help to discriminate the texts. Hence, we remove these corpus-specific stopwords from the vocabulary. Then, we convert the texts d_i to a vector representation $v_i \in \mathbb{R}^q$ according to the VSM [SWY75], where q is the reduced number of terms. Finally, we denote the vectorized dataset with $C = \{(v_i,t_i)\}_{i=1}^{n}$.

3.2. Linear SOM-based Clustering of Texts

Given C, we seek a 1D projection that projects similar texts to similar 1D positions. Therefore, we seek a sequence $S = (s_1,\ldots,s_k)$ of cells $s_i \in \mathbb{R}^q$ that adapts to texts in the original space. In the SOM algorithm the radius parameter controls how an update affects neighboring cells. Therefore, the sequence of cells model an incremental transition across the cells. This property lets us visualize a gradual change of themes across the vertical. We fit a 1D SOM to the data (as depicted in Figure 2a). We train the SOM with k cells for T steps using an initial radius of r (i.e., the distance from which we do not update cells anymore). We decrease the radius r linearly to zero, the learning rate decreases linearly from 0.5 to zero, and we use the Euclidean norm $\|\cdot\|$ as a distance measure. For training, we employ a sparse SOM implementation in combination with sparse matrices. Figure 2a exemplifies how a linear SOM adapts to a dataset.

For clustering data, the central question is how to choose the number of clusters. The optimal number depends on the dataset, which we need to analyze before clustering. Such analysis methods, however, require the repeated execution of clustering itself. Methods such as the elbow or the jump method [SJ03] are not feasible because these methods take longer than the final clustering. Therefore, we draw back to this heuristic: we set the number of clusters to $k = \lceil 2 \cdot \sqrt{n} \rceil$. In previous works, the number of training steps is in the order of hundreds per cells [Koh00]. Therefore, we set $T = 100k$ and $r = 0.1k$ (i.e., in the beginning, an update influences 10% of the neighboring cells).

3.3. UMAP-based Projection of Texts

Given the vectorized dataset C, we seek a 1D projection that preserves the distances of the original space in the projected space. As an alternative to the SOM, we can achieve this directly with UMAP. UMAP directly maps texts, represented as vectors, to a scalar value p with a 1D projection. Then, we normalize the projected values to $[0,1]$. We use the implementation by McInnes, Healy, Saul, and Grossberger [MHSG18] with the number of components set to one. For this work, it is unproblematic that UMAP maps elements to the same locations, and therefore, we set the minimum distance to zero. As the SOM updates a cell by adding the weighted observation vector, we cannot use the cosine similarity to compute the weight and, therefore, use the Euclidean distance. In contrast, UMAP computes a neighbor graph, for which we can use the cosine similarity. For the remaining parameters, we use the default values.

The result of the SOM clustering is a sequence of k cells that represent average text vectors. From these vectors, we extract terms to generate a thematic description. However, UMAP does not provide such cells. Therefore, we create k equidistant points in the projected space that cover the entire range and inversely project them to get a comparable sequence. After computing a well-generalized model for any of the proposed methods, we can project an arbitrary number of texts because training and mapping are uncoupled.

3.4. Mapping of Texts

In the mapping step, we generate the diagram. We seek to position each text in a 2D diagram (Figure 2b). The UMAP-based method provides us directly with a position p. For the SOM-based method, we need to compute such a position. Given a document representation v, we calculate its normalized vertical position $p \in [0,1]$: Let l be the total length of the sequence of cells $l = \sum_{i=2}^{k} \|s_{i-1} - s_i\|$ and b the best matching cell of v: $b = \arg\min_j \|v - s_j\|$. Then, the vertical position of document vector v is its relative position on the sequence of cells $p = (\sum_{i=2}^{b} \|s_{i-1} - s_i\|)/l$. With this information, we can count the text intensity at each location in an image represented by matrix $M \in \mathbb{R}^{k \times t}$. For example a text that is projected to the vertical position k and appears at time t increases the intensity in $M_{k,t}$ by one. Doing this for all texts yields overall patterns: $M_{k,t} = |\{(v_i, t_i) \in C | k = f(v_i) \wedge t = g(t_i)\}|$. Here f maps texts to $[1,h] \cap \mathbb{N}$ according to the previous normalization and g maps points in time linearly to $[1,w] \cap \mathbb{N}$.

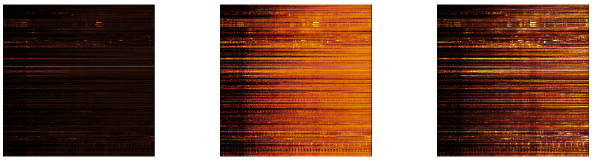
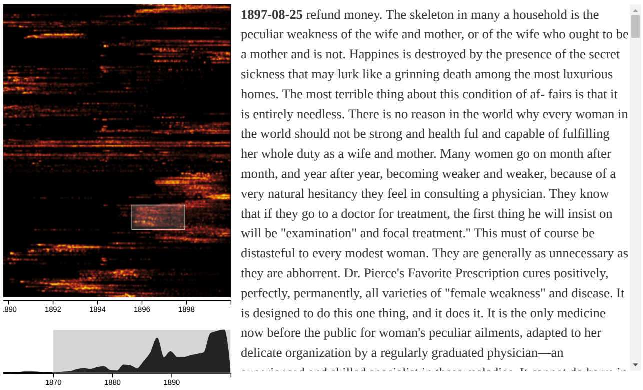

(a) *Linear normalization of M's values* **(b)** *Logarithmic normalization of M's values.* **(c)** *Column-individual rank normalization of M's values.*

Figure 3: *For creating the images, we experimented with different normalizations. Given a matrix M in which each cell denotes the frequency of texts, we seek to normalize these frequencies to make patterns visible. Mapping the frequencies linearly to a color value results in (almost) black images (Figure a). To reduce the range of the frequencies in M, we take the logarithm of all values, which reveals more patterns (Figure b). Normalizing the columns independently reveals even more patterns. Figure (c) shows an example in which we replaced the values of each column of M by their ranks.*

Computing M for various datasets, we realize that the range of values is high. Mapping the densities linearly to a color value yields images such as in Figure 3a or Figure 3b. We find that scaling columns independently and replacing each density by their rank across the columns shows patterns more distinctly than in the previous examples (e.g., Figure 3c). We make the densities visible by using a color scale inspired by the black-body radiation (see [Mor16] for color scale examples and the full color definition).

3.5. Prototype

To create a useful system for analyzing text collections, we create a prototypical system. This system allows analysts to explore a text collection and to retrieve texts based on the diagram. Figure 1 shows the system's user interface (UI) with an exemplary text collection. The UI in Figure 1 consists of the temporal-thematic diagram a, a list view of important terms b, a line chart depicting the number of texts over time c, and labels to highlight prominent patterns d. Analysts can interact with the interface in the following ways: they can point to a location of the diagram (e.g., region e). By pointing to locations of interest, the interface shows the terms of the respective theme. The interface shows as many terms as there is screen space on the left. Analysts can filter the data by selecting a time range (line chart below the diagram), which regenerates the diagram for that range. On the right, the visualization shows heuristically generated labels. The heuristic selects salient vertical positions spread across the entire vertical axis. Analysts can delete default labels and create custom labels at any



1897-08-25 refund money. The skeleton in many a household is the peculiar weakness of the wife and mother, or of the wife who ought to be a mother and is not. Happines is destroyed by the presence of the secret sickness that may lurk like a grinning death among the most luxurious homes. The most terrible thing about this condition of af- fairs is that it is entirely needless. There is no reason in the world why every woman in the world should not be strong and health ful and capable of fulfilling her whole duty as a wife and mother. Many women go on month after month, and year after year, becoming weaker and weaker, because of a very natural hesitancy they feel in consulting a physician. They know that if they go to a doctor for treatment, the first thing he will insist on will be "examination" and focal treatment." This must of course be distasteful to every modest woman. They are generally as unnecessary as they are abhorrent. Dr. Pierce's Favorite Prescription cures positively, perfectly, permanently, all varieties of "female weakness" and disease. It is designed to do this one thing, and it does it. It is the only medicine now before the public for woman's peculiar ailments, adapted to her delicate organization by a regularly graduated physician—an

Figure 4: *Example of the UI showing the list of texts corresponding to the region in the diagram which the analyst selected to inspect.*

position and name them. They can also select a rectangular region in the diagram. This hides the labels on the right and shows the corresponding texts (Figure 4). For this, the system translates the rectangular region to a range query for the specified time and cluster range on the SOM. In case of the UMAP method, the system translates the query to the specified time and the 1D embedding. Then, the system lists all texts that match the query on the right. If there is no rectangular selection in the image, the system shows the labels (see Figure 1(d)). The idea of the labels is to point to and characterize interesting vertical locations. The number of these labels is limited by the available space in this view and the labels are selected based on the luminance in the image. We implemented this prototype using a client-server model. The client uses the D3 library [BOH11], which provides low-level modules for the UI. It retrieves the diagram as an image and puts it in a Hypertext Markup Language (HTML) canvas element.

4. Evaluation

The objective of this work is to detect yet unknown temporal-thematic patterns in texts. We validate the diagram technique using qualitative result inspection (QRI) [IIC*13] and conduct semi-structured interviews with domain experts. The rationale for the first validation is to determine whether the patterns visible in the diagram are meaningful. There are no established benchmark corpora with temporal text collections and labeled texts. Therefore, we experiment with real-world datasets and qualitatively discuss the results. The technical utility of our approach will be demonstrated by linking patterns in the visualization to meaningful patterns in data. However, this validation does not answer whether domain experts can utilize the diagram as part of a system. Specifically, we would like to know if domain experts understand how to read the diagram and extract knowledge by interacting with a prototypical system. Hence, in the second validation step, we interview domain experts in a semi-structured way.

4.1. Experiments

In the following, we report on experiments with six datasets. Table 1 lists experimental parameters with the SOM and the UMAP method. To make both methods comparable, we set the parameters for text processing in the same way (especially $f_{\min}$ and $f_{\max}$). For the SOM method, we processed all the texts in the datasets. During computation, UMAP creates data structures such as a neighbor graph, whereas the SOM only updates the clusters. For determining temporal-thematic clusters with UMAP, we processed at most $m = 200{,}000$ texts due to memory limitations. The resulting diagrams in Table 3 show three of the six text collections.

4.2. Qualitative Result Inspection

Table 3 shows the application of our approach to three different text collections using the UMAP-based and the SOM-based projection procedure, respectively. We added exemplary turquoise annotations to the diagrams shown in Table 3. The central question is whether patterns are meaningful. Technically, a dense region shows an increase in the number of texts for a specific combination of terms. It is unlikely that dense regions show unrelated texts because the methods try to project similar texts to similar vertical positions. We can exemplify this argument using the results from the first experiment (first row in Table 3). This is a general collection of 300,000 newspaper articles from the LIBRARY OF CONGRESS [Lib21]. Although there is no ground truth dataset for our work, we can compare patterns with common knowledge. The diagram shows the American Civil War denoted by s and elections denoted by repeating patterns r. For example, we see the terms “soldier, army, officer, sailor” for the vertical position of s in the SOM method and “insurgent, corp, boot, colonel, military” for this label in the UMAP method. However, some areas appear to be noisy. By looking at static result images, the viewer cannot distinguish whether there are no patterns in the data or the method failed to detect them. Table 2 shows a cutout of the terms associated with the vertical position marked by c (we provide the full lists in the supplemental materials). From the gradual change and coherence of terms, we conclude that this is likely due to a lack of patterns in the data.

The second row of results is a collection of law cases [The21] from the United States of America. In the images of both variants, we can see patterns for which we added exemplary annotations. In the result from the UMAP variant, there are various patterns (e.g., labels g and f). However, we were not able to extract a concise meaning by looking at the labels created for the respective regions. One can extract meaning by looking at the labels that we created for the regions. The results from the SOM variant show patterns (e.g., labels h and i). In contrast, we find explanatory patterns, which one can see by looking at the coherence of terms within a label and across labels nearby.

The third experiment is a specific collection about advertisements Dr. Pierce’s Favorite Prescription. Pierce was a doctor who manufactured and sold quack medicine. To create a diagram about this collection, we extracted nineteenth-century newspaper articles that contain the terms *prescription* and *pierce* from the LIBRARY OF CONGRESS [Lib21]. As these texts contain optical character recognition (OCR) errors, we allowed a Levenshtein distance of one for the two terms. Because of this extraction, the collection contains similar texts with subtle differences. By looking at the results from the UMAP-based projection, we see patterns that look like random noise. This is an indication that SOM-based projection

Table 1: *Datasets and parameters (as defined in Section 3) used for the experiments. The columns D. * show the training duration and M. * show the memory consumption of the SOM and UMAP method respectively. An Intel® Core™ i5-9600K CPU computed the experiments.*

No.	Dataset	n	m	q	f_{min}	f_{max}	k	T	r	D. SOM	M. SOM	D.UMAP	M. UMAP
1	General	287,517	150,000	15,454	100	0.5	1,500	150,000	60	233 h	15 GiB	19 h	9.10 GiB
2	Law Cases	322,286	300,000	59,091	20	0.85	1,100	110,000	100	188 h	1 GiB	14 h	15.80 GiB
3	Prescription	217,141	200,000	58,694	20	0.90	932	95,000	100	275 h	0.93 GiB	23 h	11.30 GiB
4	Financial	612,486	500,000	11,676	40	0.9	1,200	30,000	120	104 h	1.78 GiB	15 h	9 GiB
5	Arxiv	41,000	41,000	12,376	5	0.9	405	40,500	40	4 h	1.42 GiB	2 h	1.22 GiB
6	Tweets	1,000,000	200,000	6,669	100	0.9	2,000	200,000	200	9 h	1.40 GiB	5 h	7.40 GiB

Table 2: *Cutout of terms from the General dataset. The region corresponding to this cutout has the label c in Table 3 row one.*

book illustrated history complete hutchinson contain
book books bible school geography street history soldiers
books book stationery street binding lectures printing camp
books stationery book price paper hindley slater variety
books games agents edition memoir cards pledge corner
brick story bath s4 near term modern easy
story tell life book american novel true window
story serial magazine series paper short write stories
story author serial paper tale poem illustrated romance
author story reader life book character original stables
author book life read romance reader title fortune
author book american article various literary essay publish

uncovers more patterns than the UMAP-based one. In combination with the terms on the left, it is unlikely that an analyst can extract useful information. The result from the SOM-based projection, however, shows various patterns. In the diagram, one can see that the collection consists globally of three types of texts (upper patterns, middle patterns, lower patterns). Intuitively, the patterns in the upper half depict individual accumulations of texts (e.g., v). Looking at the corresponding labels, one can see a variety of terms partly referring to medical topics. In the lower half, there is a progression of patterns (e.g., t). This indicates similar and changing textual content, which is supported by looking at the labels in the lower half.

The results of the remaining three experiments (financial news [Kag19], Arxiv abstracts [Kag20], random sample of tweets [Twi20]) do not show patterns (similarly to the results from the first experiment). Due to space limitations, we show the results in the supplemental material only. There are mainly two reasons for such results: There are no salient patterns in a text collection (e.g., a collection of random texts); or our method failed to detect patterns. Regarding the financial news dataset and the Arxiv dataset, we think the first reason is more likely. This can be seen by looking at the term lists. For example, the first forty rows of the Arxiv dataset show terms about NLP research that gradually changes to terms about topic modeling and recommender systems. Similarly, the first thirty rows of the financial news dataset show terms about monetary policy and currency that gradually changes to stock market-related terms. In contrast, the term list of the Twitter dataset does not show such gradual changes, and individual rows seem like a random combination of terms. Therefore, we conclude that our method cannot capture patterns for a thematically broad collection of short texts such as tweets.

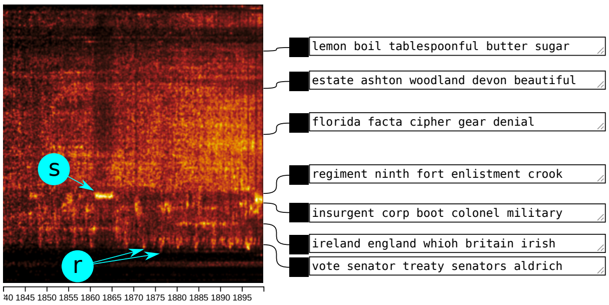
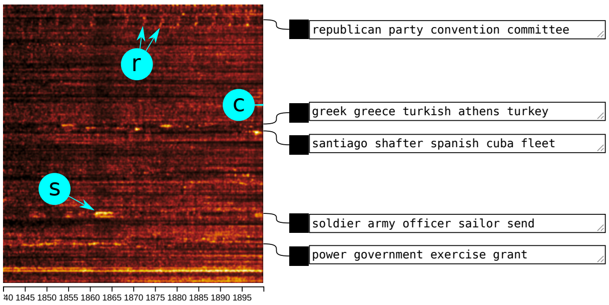
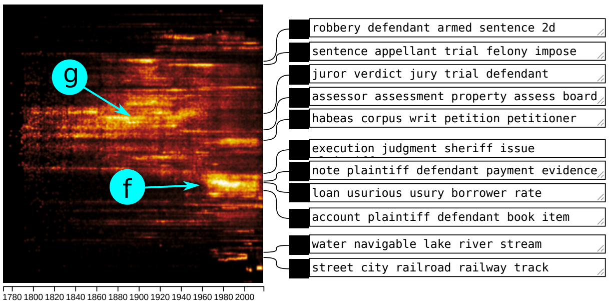
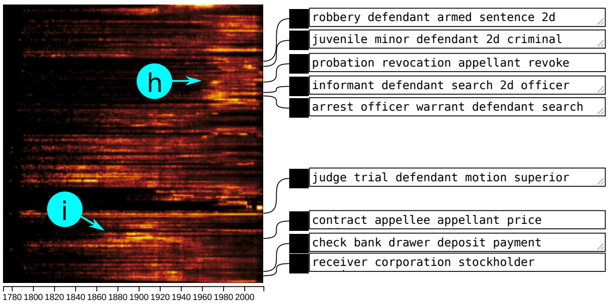
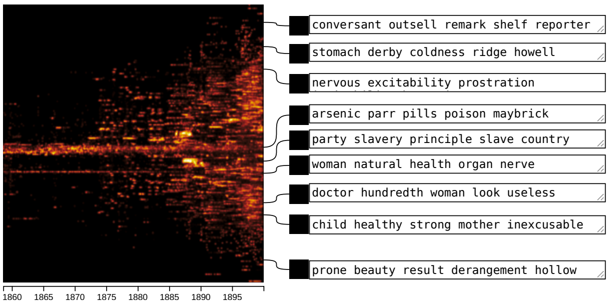
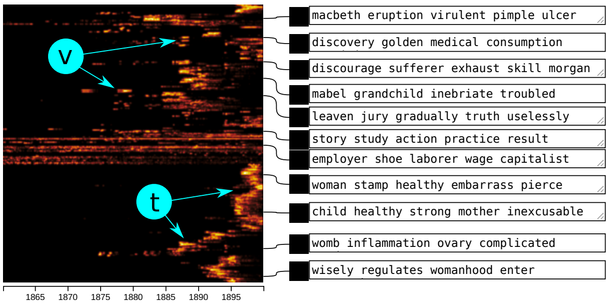
To summarize the result inspection, we find that our method can show patterns for specific types of text collections. Overall, these results indicate that the method works for collections containing long texts (hundreds of words) with a limited thematic variety (in contrast to a general sample of texts). For these experiments, the UMAP variant shows less distinctive patterns.

4.3. Semi-Structured Interviews with Historians

To test and find out if domain experts understand how to read the diagram and extract knowledge by interacting with our prototypical system, we conducted semi-structured interviews. Semi-structured interviews allow us to test planned tasks and learn about the unexpected. A classical study that measures time or errors is not adequate in this case because it cannot capture qualitative explanations from the participants. For each interview, we planned for one hour. Before conducting the interview, the experimenter (the first author of this work) informed the interviewees about the purpose of the interview. Then, the experimenter introduced the diagram and the prototypical system and told the interviewees that they could ask questions at any time or come up with their ideas. The interviews were conducted in an online session with mutual screen-sharing to explain the system and see the interviewee working with the system. The system is running as a web application, and the interviewees were encouraged to test it beforehand. For both interviews, we used the Prescription dataset.

We interviewed two persons who deal with texts collections as digital or cultural historians (a research fellow and a post-doctoral researcher). After talking about introductory information, the experimenter explained how to read the diagram and how to use the different features for exploring the dataset. The experimenter asked what it means if patterns appear at the same height or change the vertical position over time. The interviewees answered this question without hesitation or even stated this relationship on their own. The cultural historian also described the patterns t in the third row of Table 3 as “subtle development of related texts.” This indicates that the diagram is intelligible to historians outside the domain of visualization. The experimenter asked about the ways one can interact with the visualization. To explore the data, the interviewees inspected the term list on the left and created rectangles above areas of interest. To the experimenter, it was apparent that the inter-

Table 3: *Resulting images from the experiments. The left column shows the results of experiments with the UMAP method. The right column shows the results of experiments with the SOM method. The rows correspond to the datasets General, Law Cases, and Prescriptions. We created the annotations in the images to simplify the discussion of results.*

No.	UMAP Method	SOM Method
1		
2		
3		

viewees understood the meaning of the interactive features. However, the interviewees mostly reported general observations about the data set shown and they saw the global composition of the diagram (e.g., they stated that the diagram of the Prescription dataset consists of three major parts). Both interviewees mentioned that they like the prototype, and one specified that the "map metaphor" suits her thinking.

Summarizing this discussion, we find that historians can understand the concept and interact with the prototype. During the hour-long interview, it became clear that the historians could work with the prototype and could to examine interesting patterns. In the next step, we proposed to use our diagram with a dataset chosen by them to explore which types of patterns they can find.

5. Discussion and Conclusion

Our method transforms a temporal text collection into a 2D diagram that shows patterns in the collection. Experimenting with real-world text collections, we find meaningful patterns in the diagrams. This finding was not surprising because 2D projections with scatter plots have been applied to text collections. However, the experimental findings indicate that for thematically narrow collections with long texts, this method works better than for broad collections. Additionally, the experiments do not provide evidence for collections with short texts (e.g., abstracts or microblog messages). Finally, the findings from two interviews with historians indicate that 1D projections in a visualization system allow exploring the content of text collections.

Acknowledgments

This work has been funded by the German Research Foundation DFG in the context of the Digging into Data Challenge project Oceanic Exchanges and in the context of the MicroVis project in cooperation with the French National Research Agency (ANR).

References

[BJC*19] BUCHMÜLLER, J., JÄCKLE, D., CAKMAK, E., et al. "MotionRugs: Visualizing Collective Trends in Space and Time". *IEEE Trans. Vis. Comput. Graph.* 25.1 (Jan. 2019), 76–86 3.

[BNJ03] BLEI, DAVID M, NG, ANDREW Y, and JORDAN, MICHAEL I. "Latent dirichlet allocation". *J. Mach. Learn. Res.* 3.1 (2003), 993–1022 2.

[BOH11] BOSTOCK, MICHAEL, OGIEVETSKY, VADIM, and HEER, JEFFREY. "D3 Data-Driven Documents". *IEEE Trans. Vis. Comput. Graph.* 17.12 (Dec. 2011), 2301–2309 5.

[BSH*16] BACH, B., SHI, C., HEULOT, N., et al. "Time Curves: Folding Time to Visualize Patterns of Temporal Evolution in Data". *IEEE Trans. Vis. Comput. Graph.* 22.1 (Jan. 2016), 559–568 2.

[CC00] COX, TREVOR F and COX, MICHAEL AA. *Multidimensional Scaling*. Chapman and Hall, 2000 2.

[CLT*11] CUI, WEIWEI, LIU, SHIXIA, TAN, LI, et al. "TextFlow: Towards Better Understanding of Evolving Topics in Text". *IEEE Trans. Vis. Comput. Graph.* 17.12 (2011), 2412–2421 3.

[CM84] CLEVELAND, WILLIAM S. and MCGILL, ROBERT. "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods". *J. Am. Stat. Assoc.* 79.387 (1984), 531–554 2.

[CVW09] COLLINS, C., VIEGAS, F. B., and WATTENBERG, M. "Parallel Tag Clouds to explore and analyze faceted text corpora". *IEEE VAST*. 2009 3.

[DYW*13] DOU, WENWEN, YU, LI, WANG, XIAOYU, et al. "Hierarchicaltopics: Visually exploring large text collections using topic hierarchies". *IEEE Trans. Vis. Comput. Graph.* 19.12 (2013), 2002–2011 2.

[ESD*18] EL-ASSADY, MENNATALLAH, SPERRLE, FABIAN, DEUSSEN, OLIVER, et al. "Visual analytics for topic model optimization based on user-steerable speculative execution". *IEEE Trans. Vis. Comput. Graph.* 25.1 (2018), 374–384 3.

[HHWN02] HAVRE, S., HETZLER, E., WHITNEY, P., and NOWELL, L. "ThemeRiver: visualizing thematic changes in large document collections". *IEEE Trans. Vis. Comput. Graph.* 8.1 (2002), 9–20 2.

[IIC*13] ISENBERG, T., ISENBERG, P., CHEN, J., et al. "A Systematic Review on the Practice of Evaluating Visualization". *IEEE Trans. Vis. Comput. Graph.* 19.12 (2013), 2818–2827 5.

[JFSK16] JÄCKLE, D., FISCHER, F., SCHRECK, T., and KEIM, D. A. "Temporal MDS Plots for Analysis of Multivariate Data". *IEEE Trans. Vis. Comput. Graph.* 22.1 (2016), 141–150 3.

[Jol11] JOLLIFFE, IAN. *Principal Component Analysis*. Springer, 2011 2.

[JS98] JERDING, D. F. and STASKO, J. T. "The Information Mural: a technique for displaying and navigating large information spaces". *IEEE Trans. Vis. Comput. Graph.* 4.3 (1998), 257–271 3.

[Kag19] KAGGLE INC. *US Financial News Articles*. 2019. URL: https://www.kaggle.com/jeet2016/us-financial-news-articles (visited on 03/09/2021) 6.

[Kag20] KAGGLE INC. *Arxiv Papers Metadata Dataset*. 2020. URL: https://www.kaggle.com/tayorm/arxiv-papers-metadata (visited on 02/15/2020) 6.

[Kei00] KEIM, D. A. "Designing pixel-oriented visualization techniques: theory and applications". *IEEE Trans. Vis. Comput. Graph.* 6.1 (2000), 59–78 3.

[KKE20] KNITTEL, J., KOCH, S., and ERTL, T. "PyramidTags: Context-, Time- and Word Order-Aware Tag Maps to Explore Large Document Collections". *IEEE Trans. Vis. Comput. Graph.* (2020), 1–1 3.

[Koh00] KOHONEN, TEUVO. "Self-Organizing Maps of Massive Document Collections". *IJCNN*. IEEE Computer Society, 2000, 3–12 2, 4.

[Lib21] LIBRARY OF CONGRESS. *Chronicling America: Historic American Newspapers*. 2021. URL: https://chroniclingamerica.loc.gov/ (visited on 03/05/2021) 5.

[LYK*12] LUO, D., YANG, J., KRSTAJIC, M., et al. "EventRiver: Visually Exploring Text Collections with Temporal References". *IEEE Trans. Vis. Comput. Graph.* 18.1 (2012), 93–105 2.

[MHSG18] MCINNES, LELAND, HEALY, JOHN, SAUL, NATHANIEL, and GROSSBERGER, LUKAS. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction". *J. Open Source Softw.* 3.29 (2018), 861 4.

[Mor16] MORELAND, KENNETH. "Why we use bad color maps and what you can do about it". *HVEI* 2016.16 (2016), 1–6 4.

[RAK10] RZESZUTEK, RICHARD, ANDROUTSOS, DIMITRIOS, and KYAN, MATTHEW. "Self-Organizing Maps for Topic Trend Discovery". *IEEE Signal Process. Lett.* 17.6 (2010), 607–610 2.

[RSB*17] RUPPERT, TOBIAS, STAAB, MICHAEL, BANNACH, ANDREAS, et al. "Visual Interactive Creation and Validation of Text Clustering Workflows to Explore Document Collections". *Visualization and Data Analysis 2017*. Ed. by WISCHGOLL, THOMAS, ZHANG, SONG, and KAO, DAVID L. 2017, 46–57 3.

[Sar13] SARLIN, PETER. "Self-organizing time map: An abstraction of temporal multivariate patterns". *Neurocomputing* 99 (2013), 496–508 2, 3.

[SBLK09] SCHRECK, TOBIAS, BERNARD, JÜRGEN, LANDESBERGER, TATIANA VON, and KOHLHAMMER, JÖRN. "Visual Cluster Analysis of Trajectory Data with Interactive Kohonen Maps". *Inf. Vis.* 8.1 (2009), 14–29 2.

[SJ03] SUGAR, CATHERINE A and JAMES, GARETH M. "Finding the number of clusters in a dataset: An information-theoretic approach". *J. Am. Stat. Assoc.* 98.463 (2003), 750–763 4.

[SWY75] SALTON, G., WONG, A., and YANG, C. S. "A vector space model for automatic indexing". *Commun. ACM* 18.11 (1975), 613–620 2, 3.

[The21] THE PRESIDENT AND FELLOWS OF HARVARD UNIVERSITY. *Caselaw Access Project*. 2021. URL: https://case.law/bulk/download/ (visited on 03/04/2021) 5.

[Twi20] TWITTER, INC. *Docs*. 2020. URL: https://developer.twitter.com/en/docs (visited on 04/03/2020) 6.

[vdMH08] Van der MAATEN, LAURENS and HINTON, GEOFFREY. "Visualizing Data using t-SNE". *J. Mach. Learn. Res.* 9.11 (2008), 2579–2605 2.

[VWD04] VIÉGAS, FERNANDA B., WATTENBERG, MARTIN, and DAVE, KUSHAL. "Studying Cooperation and Conflict Between Authors with History Flow Visualizations". ACM, 2004, 575–582 2.

[WLWK08] WU, HO CHUNG, LUK, ROBERT WING PONG, WONG, KAM FAI, and KWOK, KUI LAM. "Interpreting TF-IDF term weights as making relevance decisions". *ACM Trans. Inf. Syst.* 26.3 (2008), 1–37 3.

[WV08] WATTENBERG, MARTIN and VIÉGAS, FERNANDA B. "The word tree, an interactive visual concordance". *IEEE Trans. Vis. Comput. Graph.* 14.6 (2008), 1221–1228 3.

[XLG03] XU, WEI, LIU, XIN, and GONG, YIHONG. "Document clustering based on non-negative matrix factorization". *SIGIR*. 2003 2.