
Otto-von-Guericke University Magdeburg



Department of Computer Science
Institute of Simulation and Graphics

Master Thesis

Pattern detection in tabular data with shallow hierarchy: A Visual Analytics case-study in Narrative Visualisation

Author:

Jalaj Arjav, Vora
221510

June 19, 2023

Supervisors:

First Supervisor

Prof. Dr.-Ing. habil. Bernhard Preim

Department of Computer Science
Otto-von-Guericke University
Universitätsplatz 2
39106 Magdeburg, Germany

Second Supervisor

Prof. Dr.-Ing. habil. Holger Theisel

Department of Computer Science
Otto-von-Guericke University
Universitätsplatz 2
39106 Magdeburg, Germany

Vora, Jalaj Arjav:

*Pattern detection in tabular data with shallow hierarchy: A Visual Analytics
case-study in Narrative Visualisation*

Master Thesis, Otto-von-Guericke University
Magdeburg, 2023.

Contents

Abstract

1 Introduction and Motivation

1.1 Motivation	1
1.2 Contribution	2
1.3 Structure of this thesis	3

2 Theoretical Background

2.1 Narrative Visualization, Storytelling Design Space	5
2.1.1 Narrative Visualization	5
2.1.2 Storytelling Design Space	5
2.2 Visual Analytics	8
2.2.1 Components of Visual Analytics	8
2.2.2 Framework for Visual Analytics	9
2.3 Dimensionality Reduction	9
2.3.1 t-Distributed Stochastic Neighbor Embedding	10
2.3.2 t-SNE based on Barnes-Hut Approximation	11
2.4 Bi-Clustering	13
2.4.1 BiMax	13

3 Related Work

3.1 Visual Analytics applications in General	15
3.2 Visual Analytics application using tabular data	15

4 Design and Implementation

4.1 Overview of Architecture	17
4.1.1 Software	21
4.2 Interactive Visualizations: t-SNE	22
4.2.1 General Overview: t-SNE projections	22
4.2.2 Filtering and re-computation	24
4.2.3 Deviation of Subspace plot	26
4.2.4 Usage deviation plot	27
4.3 Interactive Visualizations: Bi-Clustering	28
4.3.1 General Overview: Bi-Clusters	28
4.3.2 Filtering and re-computation	29

4.3.3	Dendrogram Plot	30
4.3.4	Heatmap	30
5	Evaluation	
5.1	Evaluation Overview	33
5.2	Evaluation Set-up	33
5.3	Evaluation Discussion	35
5.3.1	Analysis of User-studies	35
5.4	Limitations	39
6	Conclusions and Future Work	
6.1	Conclusions	41
6.2	Future Work	41
A	Abbreviations and Notations	
B	List of Figures	
C	List of Tables	
D	Bibliography	

Abstract

This master's thesis explores the application of visual analytics to detect patterns in tabular data with a shallow hierarchy. The study focuses on developing a novel approach that combines data exploration, automated data analysis, and interactive visualizations to facilitate pattern recognition and understanding in storytelling design space. The research methodology involved a user-study with four participants where the application was evaluated, utilizing the storytelling design space dataset.

The results demonstrate the effectiveness of the proposed approach in detecting patterns within tabular data with shallow hierarchical structures. The interactive visualizations enabled users to explore the data at multiple levels of detail, revealing underlying relationships and insights.

The findings suggest that this approach holds promise for usability. The implications of the research include the development of more effective tools and techniques for data analysts, researchers, and domain experts to explore and uncover insights in storytelling design space or tabular data with shallow hierarchy following similar structure and properties with storytelling design space.

"If you can't explain it simply, then you don't understand it well enough"

Albert Einstein

Acknowledgements

I would like to thank *Prof. Dr.-Ing. habil. Bernhard Preim* and *Prof. Dr.-Ing. habil. Holger Theisel* for their supervision and support. This Thesis would have been impossible without incredible support, advice, guidance and patience of my advisor *M.Sc. Benedikt Mayer*. I would also like to thank *Dr.-Ing. Monique Meuschke, Sarah Mittenentzwei and Anna Kleinau* for voluntary participation and their invaluable contribution of insights in evaluation.

I would also like to thank my Guru *Kamlesh Vyas* for his continuous support, motivation and blessings. I would also like to thank *Dharmin Bakraniya* for extremely fruitful discussions, ideas and suggestions on implementation. Lastly, I would like to thank my friends and my family for continuous love and support.

1

Introduction and Motivation

1.1 Motivation

The importance of data in today's digital age cannot be overstated. Data is becoming increasingly important and is also considered new oil. With the exponential growth of information and technology, organizations and individuals alike are generating vast amounts of data. This wealth of data holds valuable insights and patterns waiting to be discovered.

However, real-world data is commonly characterized by its large size and complexity. Therefore to structure such data in a tabular format, resembling database tables or spreadsheets contains multiple attributes or dimensions would be one of the intuitive idea.

Tabular data has also emerged as a fundamental method for managing data and has gained significant adoption across various fields and by diverse individuals such as scientists, financial professionals, analysts and policymakers CHEN und CAFARELLA (2013); DOU et al. (2018); FURMANOVA et al. (2017); PERIN et al. (2014). However, such tabular data may exhibit hierarchical dependencies, where the values in certain dimensions are hierarchically organized. The hierarchical dependencies between these dimensions is crucial for uncovering hidden patterns, identifying trends, and gaining a comprehensive understanding of the underlying relationships within the dataset.

One of the examples of the such tabular data would be taxonomy or analysis of design space of visual storytelling. Such design spaces contains analysis of data curated over time. Examples of such design space would be design spaces curated by SEGEL und HEER (2010) and STOLPER et al. (2016). Here, the work focuses on finding the articles from online journalism to

see what kind of storytelling techniques do such articles use. This information is extracted and compressed in a design space such as one created by STEINHAEUER (2022). Such design spaces, contains binary information in tabular form. However, such data contains certain shallow hierarchies with fewer depth within dimensions.

Therefore, it would be great if we had a way to automatically find patterns in such tabular data with shallow hierarchies. However, often automatic algorithms depend on certain input parameters and require interactivity to understand the analysis. Therefore, this inspires to integrate the execution of the automatic algorithms with an interactive visual application that allows users to customize input parameters and visualize the analysis of output in more insightful and effective way.

Therefore, The objective of this thesis is to develop visual analytics application for storytelling design space with shallow hierarchy by leveraging interactive visualizations and advanced analytical approaches. This research aims to uncover hidden patterns, relationships, and narrative structures within the dataset. The goal is to provide a deeper understanding of the storytelling data and enhance the interpretability and storytelling capabilities of the visualizations.

1.2 Contribution

Aim of this thesis is to develop visual analytics application as an interactive tool, which helps user to explore, analyse and detect patterns from the tabular data with shallow hierarchies using advanced data analysis in conjunction with interactive visualization.

This research will contribute following:

- Develop a web-based visual analytics application for storytelling design space to detect underlying relationships and patterns
- Evaluate the developed application with user-studies for validation of research

1.3 Structure of this thesis

This thesis provides the research in following structure:

- **Section 2** provides theoretical background and understanding of topics: Visual Analytics, t-SNE and Biclustering algorithms
- **Section 3** provides brief introduction of related work
- **Section 4** explains the design designs and implementation methodology details for the development of visual analytics application
- **Section 5** describes set-up of evaluation and findings of the user-case studies with the application
- At last, **Section 6** gives conclusion of this thesis and provides future work

2

Theoretical Background

2.1 Narrative Visualization, Storytelling Design Space

2.1.1 Narrative Visualization

Narrative visualization is termed by combination of information visualization and storytelling SEGEL und HEER (2010). In visualization research community, this term is interchangeably used with Visual Story telling. This promotes narrative visualization being a new class of visualization which focuses on communicating complex data in understandable and engaging way STEINHAEUER (2022).

2.1.2 Storytelling Design Space

Storytelling design space can be defined as analysis of curated data driven stories. Works of SEGEL und HEER (2010); STOLPER et al. (2016) define and provide a curation of various storytelling techniques used by authors of online journalism against diverse stories. This work and corpus has been extended by STEINHAEUER (2022) by adding visual data stories from spatio-temporal data.

Figure 2.1 shows example of design space curated by STEINHAEUER (2022). For simplicity, this thesis uses this design space without temporal context of the data.

The selected design space consists of 130 collected stories and 35 storytelling techniques. The design space is binary in nature. Meaning it represents information such as whether a storytelling technique is present in a story or not. The design space forms a hierarchy with shallow or single-

level of depth within storytelling techniques. Out of the 35 storytelling techniques, there exists 6 Main Categories of techniques namely:

1. Genre
2. Communicating Narrative and Exploration of Data
3. Linking Separate Story Elements
4. Providing Context and Navigation
5. Providing Controlled Exploration
6. Visual Encoding of Space and Time

Each of these main categories contains various storytelling techniques as sub-categories forming a hierarchy of depth one. Design space forms binary nature of data for sub-categories representing their presence as 1 and absence as 0. Main categories follow a logical OR among all the sub-categories. Meaning if any of the sub-categories are present in the story, then the main category is automatically present.

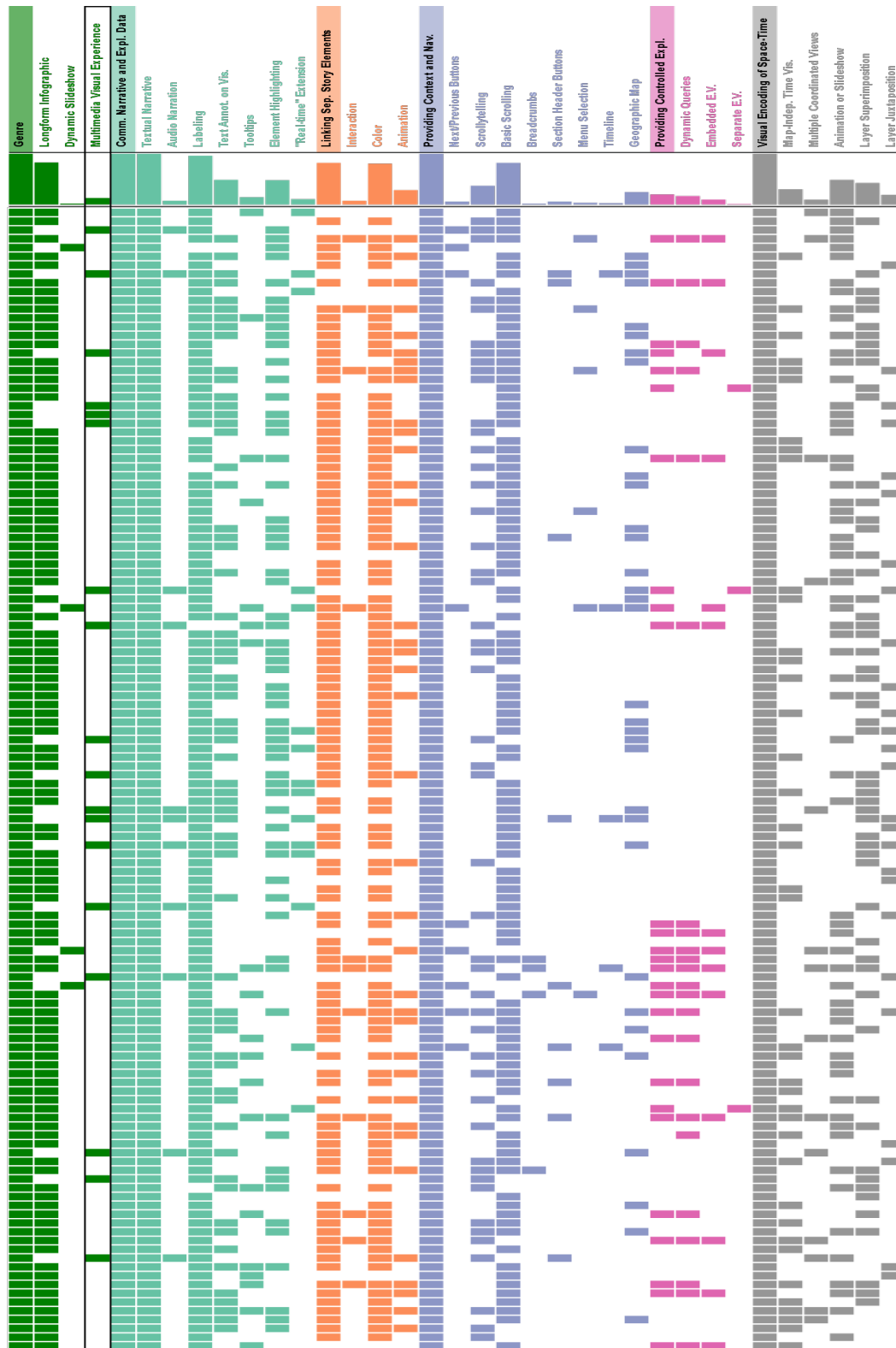


Figure 2.1: Storytelling design space analysis data curated by STEINHAEUER (2022); 130 Stories and 35 Storytelling techniques

2.2 Visual Analytics

Visual Analytics refers to the interdisciplinary field that combines interactive visualizations, data analysis, and human cognitive abilities to gain insights, discover patterns, and facilitate decision-making processes COOK und THOMAS (2005). WONG und THOMAS (2004) defined visual analytics as, "the science of analytical reasoning facilitated by interactive human-machine interfaces".

2.2.1 Components of Visual Analytics

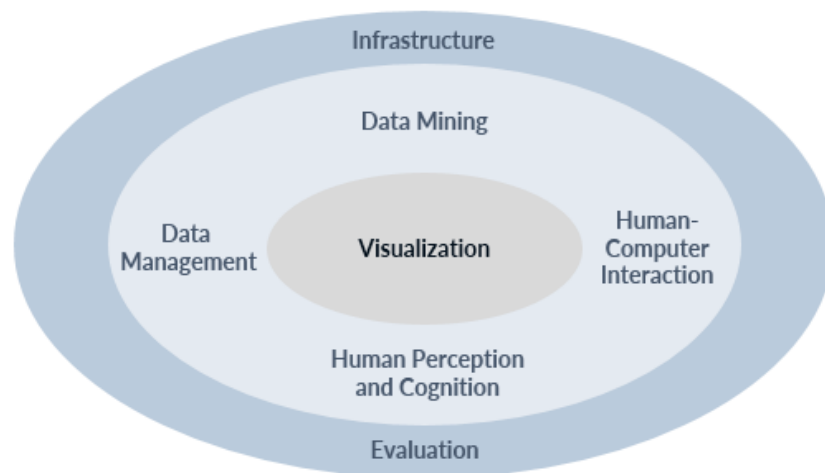


Figure 2.2: Defining Visual Analytics , combination of different disciplines COOK und THOMAS (2005)

As Figure 2.2 defines components of Visual Analytics. Figure 2.2 describes a composite architecture of interdisciplinary fields. Visualization is a central component with being crucial element between automated data analysis and human cognitive and perceptive factors. Automated Data Analysis consists of Data Mining and Data Management as automated analysis components along with Human Interactive factors such as Human Perception and Cognition and Human Computer Interaction. All of these components are bundled through a robust infrastructure and a methodological evaluation process.

The primary focus of visual analytics revolves around conducting undirected search processes to identify and elucidate trends and structures within large and complex datasets.

2.2.2 Framework for Visual Analytics

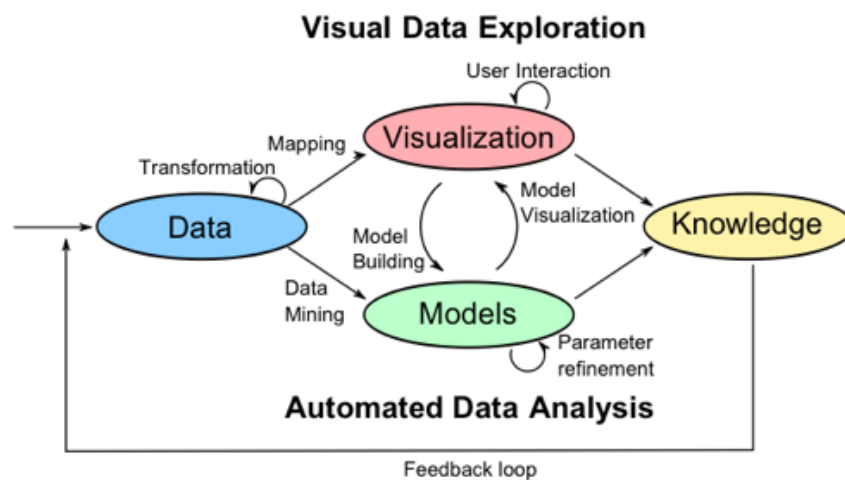


Figure 2.3: Visual Analytics process KEIM et al. (2010, 2008c)

Figure 2.3 describes visual analytics as a process. It combines automated and visual analysis methods with human interaction in order to turn data into knowledge KEIM et al. (2008b).

2.3 Dimensionality Reduction

High Dimensional data are tough to analyze and visualize with traditional knowledge discovery approaches, such as Clustering, Association Mining, Decision Trees. Such problems are tackled by Dimensionality Reduction techniques. Dimensionality Reduction Techniques involve generating lower dimensional planes with visual representation of high dimensional data by preserving structure, enabling overview of data and its distribution. Figure 2.4 follows process of Dimensionality Reduction in a Visual Analytics Process.

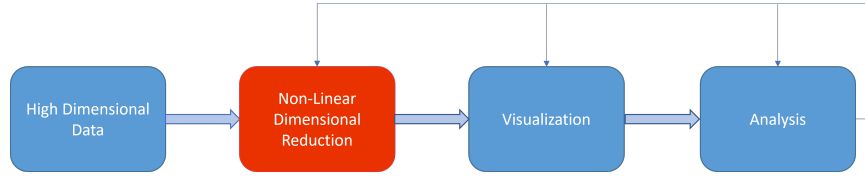


Figure 2.4: Non-Linear Dimensionality Reduction in Visual Analytics process (inspired by PREIM (2023))

For skewed or multimodal data, non-linear dimensionality reduction is used. Non-linear dimensionality reduction technique preserves small distances between the points when transforming in to lower dimensional space and has more degrees of freedom.

2.3.1 t-Distributed Stochastic Neighbor Embedding

Stochastic Neighbourhood Embedding

t-SNE is derived from SNE, proposed by HINTON und ROWEIS (2002). The goal of SNE is to preserve both global and local structure of the data when mapping from high dimensional space into 2-dimensional or 3-dimensional space. It calculates euclidean distances between points in high dimensional space and calculates their respective conditional probabilities that represent similarities. Similarity $p_{j|i}$ between x_i and x_j is represented by Equation 2.1.

$$p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / 2\sigma^2)} \quad (2.1)$$

Similarly, for projected points y_i and y_j , similarity $q_{j|i}$ is calculated as Equation 2.2

$$q_{j|i} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / 2\sigma^2)} \quad (2.2)$$

Iteratively, points are moved until respective two distributions are moved close as possible. The difference between two distributions is measured by Kullback Leibler divergence.

$$C = \sum KL(P_i \parallel Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{p_{i|j}} \quad (2.3)$$

Another crucial parameter is Perplexity i.e., variance σ_i of the Gaussian distribution. The choice of σ_i is locally adapted to the density of the probability distribution. In dense regions, smaller value of σ_i is found more appropriate than in sparser regions. The value of σ_i is computed based on Shannon Entropy with binary search.

t-Distributed Stochastic Neighbor Embedding

SNE creates effective visual representations, but it faces two challenges.

1. Firstly, optimizing its cost function is challenging.
2. Secondly, it encounters a crowding problem.

Therefore, these problems are handled by t-SNE which uses Student t-Distribution with one degree of freedom when mapping the points from high dimensional space to lower dimensional space as a heavy-tailed distribution VAN DER MAATEN und HINTON (2008).

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|y_i - y_l\|^2)^{-1}} \quad (2.4)$$

The derivative of the Kullback-Leibler divergence between probability distributions P and Q, where Q is based on the student-t distribution, can be expressed as

$$\frac{\delta C}{\delta y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)(1 + \|y_i - y_j\|^2)^{-1} \quad (2.5)$$

The Algorithm 1 is explained in detail below:

2.3.2 t-SNE based on Barnes-Hut Approximation

In the original tSNE algorithm, the force calculation is performed using a straightforward method, which leads to a computational and memory complexity of $O(n^2)$. However, Barnes-Hut-SNE (BH-SNE) VAN DER MAATEN

Algorithm 1: t-Distributed Stochastic Neighbor Embedding

Data: Dataset $X = \{x_1, x_2, \dots, x_n\}$
cost function parameters: perplexity $Perp$,
optimization parameters: number of iterations T , learning rate η ,
momentum $\alpha(t)$.
Result: low-dimensional data representation $Y^{(T)} = \{y_1, y_2, \dots, y_n\}$.

begin
 compute pairwise affinities $p_{j|i}$ with perplexity $Perp$ using Equation 2.1
 set $p_{i,j} = \frac{p_{j|i} + p_{i|j}}{2n}$
 sample initial solution $Y^{(0)} = \{y_1, y_2, \dots, y_n\}$ from $N(0, 10^{-4}I)$
 for $t = 1$ **to** T **do**
 compute low-dimensional affinities $q_{i,j}$ (using Equation 2.2)
 compute gradient $\frac{\delta C}{\delta Y}$ (using Equation 2.3)
 set $Y^{(t)} = Y^{(t-1)} + \eta \frac{\delta C}{\delta Y} + \alpha(t)(Y^{(t-1)} - Y^{(t-2)})$
 end
end

(2014) is an advanced version of tSNE that employs two specific approximations to reduce the computational complexity to $O(n \log n)$ and the memory complexity to $O(N)$.

The first approximation is based on the observation that the probability $p_{i,j}$ becomes extremely small if the data points x_i and x_j are dissimilar. Therefore, when calculating the similarities of a particular data point x_i , it is sufficient to consider only the points that are its nearest neighbors (denoted as \mathcal{N}_i). The size of \mathcal{N}_i can be determined as $K = \lfloor 3\mu \rfloor$, where μ represents the user-selected perplexity and $\lfloor \cdot \rfloor$ indicates rounding down to the nearest lower integer. While maintaining the quality of the embedding VAN DER MAATEN (2014), there is the option to utilize a sparse approximation of the high-dimensional similarities. To compute the K-Nearest Neighbors, a Vantage-Point Tree (VP-Tree) YIANILOS (1993) is employed. This data structure is capable of performing KNN queries efficiently in high-dimensional metric spaces, achieving a time complexity of $O(n \log n)$. The VP-Tree is a binary tree where each non-leaf node is associated with a hyper-sphere centered on a data-point. The left children of each node contain the points residing inside the hyper-sphere, while the right children contain the points outside of it. Since tSNE can be viewed as an N-

body simulation, the Barnes-Hut algorithm BARNES und HUT (1986) can be applied to reduce the computational complexity to $O(n \log n)$.

2.4 Bi-Clustering

Clustering is seen as a successful approach for analyzing and exploring data. Clustering algorithms aim to divide data objects into distinct clusters, optimizing the similarity within each cluster while minimizing the similarity between clusters. This is accomplished by utilizing a similarity measure to assess the resemblance between objects and form cohesive groups. The goal is to create clusters that show very high internal similarity and low similarity with objects in other clusters. Biclustering, in contrast to traditional clustering, enables the simultaneous clustering of both rows and columns in a data matrix. Rather than partitioning the objects solely along one dimension (rows or columns), biclustering identifies subsets of rows and columns that exhibit similar patterns or behaviors. This approach aims to uncover coherent substructures within the data matrix that are characterized by consistent relationships between subsets of rows and columns. By considering both dimensions, biclustering provides a more comprehensive perspective on the underlying structure and dependencies in the data BOZDAĞ et al. (2009); EREN et al. (2013).

2.4.1 BiMax

BiMax is an algorithm based on the divide and conquer approach, specifically designed to identify rectangular regions of 1's in a binary matrix PRELIĆ et al. (2006). The algorithm begins by considering the entire data matrix and recursively divides it into a checkerboard-like format. It is important to note that BiMax operates solely on binary data, so any datasets must be converted or transformed into a binary format before applying the algorithm.

3

Related Work

3.1 Visual Analytics applications in General

LI et al. (2022) have developed an interactive visual analysis tool for hierarchical tabular data constructing a model which defines row/column headings as bi-clustering and hierarchical structures to explore relationships among the hierarchical row and column labels interactively and effectively. Whereas, ECKELT et al. (2019) proposes TourDino integrated in the Ordino STREIT et al. (2019), a drug discovery platform for the purpose of identifying new drug targets. TourDino provides a supporting view that helps users, who are not experts in statistics, to verify generated hypotheses and confirm insights through exploration and validation of statistical hypotheses using interactive visualisation on tabular data. Interestingly, FURMANOVA et al. (2017) provides scalable visualisation of tabular data, providing interactive analysis through hierarchical aggregation of subsets.

3.2 Visual Analytics application using tabular data

Related work shows research in the direction of using dimensionality reduction techniques for high dimensional data to reduce it into lower dimensions combined either with supervised learning tasks such as classification or unsupervised learning tasks such as cluster analysis.

STEED et al. (2020) creates a visual exploration system for multivariate data with heterogeneous type which helps understand inputs to algorithm such as neural network. XU et al. (2020) used t-SNE algorithm as dimensionality reduction and fed the resulting low dimensional features to commonly used machine learning algorithms for compositional microbiome

data. On the other hand, NAM et al. (2007) describes the importance of cluster analysis using an interactive tool to control cluster parameters on high dimensional aerosol data.

ZHANG et al. (2021) diagnose errors and find patterns in machine maintenance log data through machine learning assisted visual analytics. Here, authors use data-type dependent dimensionality reduction technique, such as use of contrasting clusters in Principal Component Analysis (ccPCA) for numerical data, contrasting clusters in Multiple Correspondence (ccMCA) for categorical data and Uniform manifold approximation and projection (UMPA) for text data in combination with clustering. ZHOU et al. (2017) address the spatial clusters of air-quality data using visual analytics tool and exemplify factors responsible for the air-quality using MDS and Hierarchical clustering. DEVASSY und GEORGE (2020) show t-SNE outperforming PCA for forensic document analysis done on hyperspectral imaging data.

4

Design and Implementation

This section aims to elucidate the architectural framework of the developed visual analytics application. It will delve into the practical implementation of the application. This includes discussing the design decisions made for the application, as well as explaining how navigation within the system is facilitated. Additionally, it will highlight how the chosen algorithms adhere to principles in visual analytics and information visualization for pattern detection and knowledge discovery.

4.1 Overview of Architecture

The proposed application is a web-based user-centric visual analytics system that focuses on integrating clustering and subspace clustering algorithm with an interactive visualization. The implementation of the application aligns closely with a visual analytics process architecture (detail explanation in Section 2 and Figure 2.3), depicted in Figure 4.1.

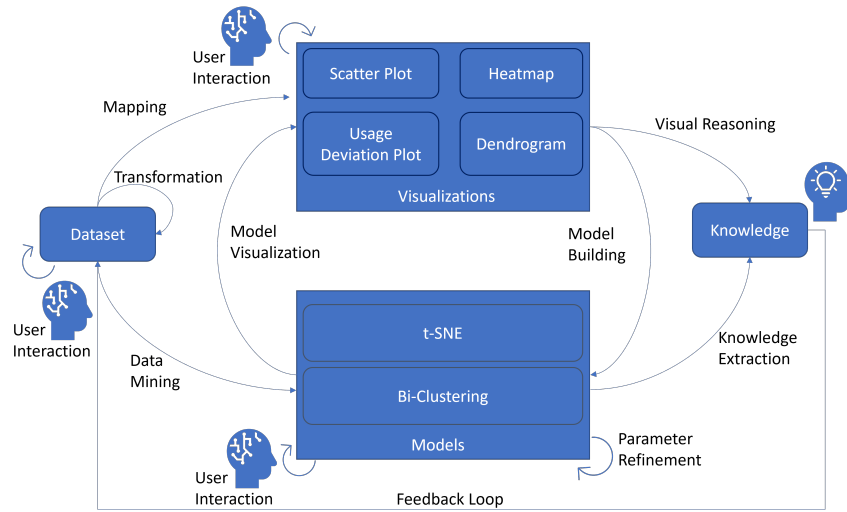


Figure 4.1: Visual Analytics Process Architecture

The application focuses on addressing following considerations related to pattern detection and relationship exploration in data:

1. reducing effort to interpret analysis result
2. preserve linkage of results
3. provide organised structure of found patterns

The fundamental concept of application entails a hierarchical exploration strategy inspired by SHNEIDERMAN (1996) mantra on visual information-seeking, which emphasizes an initial overview, followed by zooming and filtering, and finally obtaining detailed information on demand. To achieve this, the application employs various visualization widgets that are interconnected through linking-and-brushing techniques BECKER und CLEVELAND (1987); BUJA et al. (1991); KEIM (2002); SAAKE et al. (2000); VOIGT (2002). These widgets can be adjusted in terms of size and position to accommodate the users' requirements. Figure 4.2 and 4.3 describes visual representation of the application.



Figure 4.2: R-Shiny WebApp depicting different sections of the app. A) represents the drop-down menu to select a dataset. B) represents the control parameters of t-SNE algorithm; namely hierarchical normalization, number of iterations, and Perplexity as discussed in Section 2, C) represents tab to explore the algorithm, D) Scatter-plot for 2D representation of t-SNE projections, E) Scatter-plot with red line marker; the red line marker represents average pairwise similarity of the subset of data in plot D and the scatter markers represent the selected pairwise similarity of subset



Each algorithm provides two level of exploration. First, the application gives user a general overview of all detected clusters and their respective subspaces and similarity between them. It also provides information on properties of the found clusters which is visualized through hovering and interactive dendrogram plot. The overview of the similarities is provided by both t-SNE and Bi-Clustering using scatter plot. Additionally Bi-Clustering also combines scatter plot with matrix based heatmap for details.

Moving to the second exploration level, users are empowered to select a subset of relevant clusters either from the t-SNE scatter plot. Within each cluster subspace, the user has the capability to examine the distribution of cluster members across individual dimensions and make a comparative analysis between this distribution and the global distribution encompassing all data records through usage deviation plot.

4.1.1 Software

The application is constructed utilizing the R Programming language and Plotly's R graphing library¹ integration with the R-Shiny Web Application framework². R is a free and open-source³ computational tool widely utilized in various research domains, including statistics, bioinformatics, biology, physics, mathematics, chemistry, economics, geology, and medicine GIORGI et al. (2022). R offers an accessible learning curve and incorporates built-in plotting capabilities, enabling users to swiftly explore and visualize data. Moreover, R provides an extensive collection of native functions specifically designed for statistical and data science analysis, facilitating the exploration and extraction of insights from data CRAWLEY (2012); SCHMULLER (2017); WICKHAM et al. (2023). The R Shiny framework extends the capabilities of R by enabling the development of interactive and dynamic websites and web-based applications that leverage the data science functionalities of R in a consistent manner GREENE et al. (2014); JIA et al. (2022). By employing the reactive programming framework offered by R-Shiny, the process of data exploration and algorithmic analysis becomes more efficient and streamlined, thereby enhancing the knowledge discovery process through Visual Analytics SALVANESCHI et al. (2015).

¹ <https://plotly.com/r/>

² <https://jalajvora-master-thesis.shinyapps.io/Master-Thesis-Visual-Analytics/>

³ <https://www.r-project.org/>

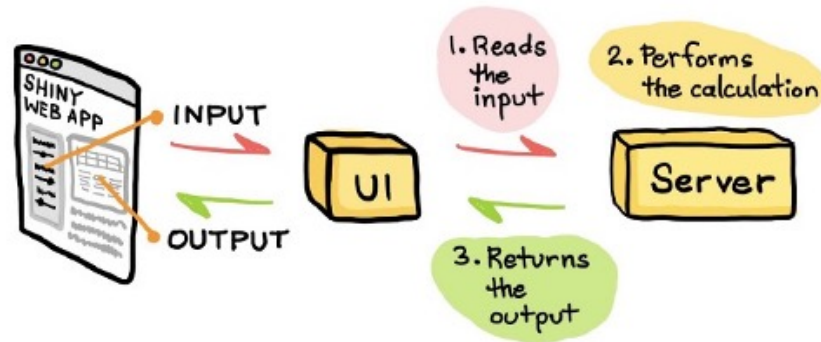


Figure 4.4: R-Shiny WebApp NANTASENAMAT (2020)

4.2 Interactive Visualizations: t-SNE

4.2.1 General Overview: t-SNE projections

The general overview of the data is given by 2-Dimensional t-SNE projections as shown in Figure 4.5 since scatter plot is commonly used visualization technique to represent t-SNE projections VENTOCILLA und RIVEIRO (2020); WAGEMANS et al. (2012). The plot represents distribution of high dimensional data into 2-dimensional space.

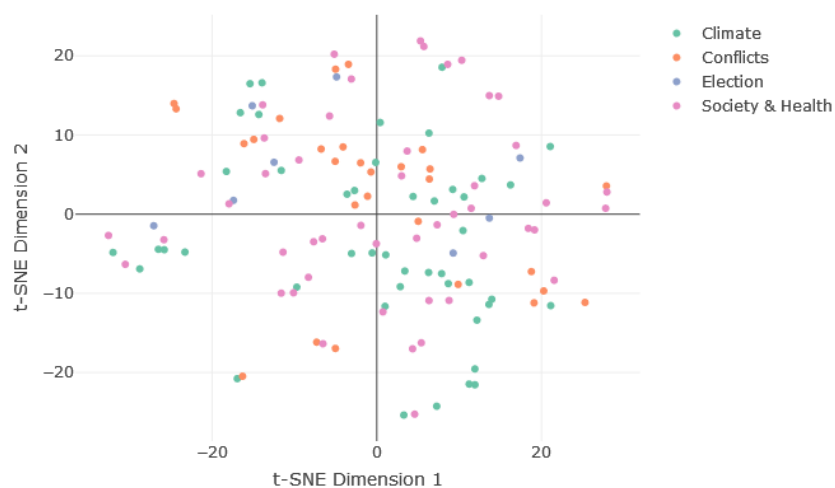
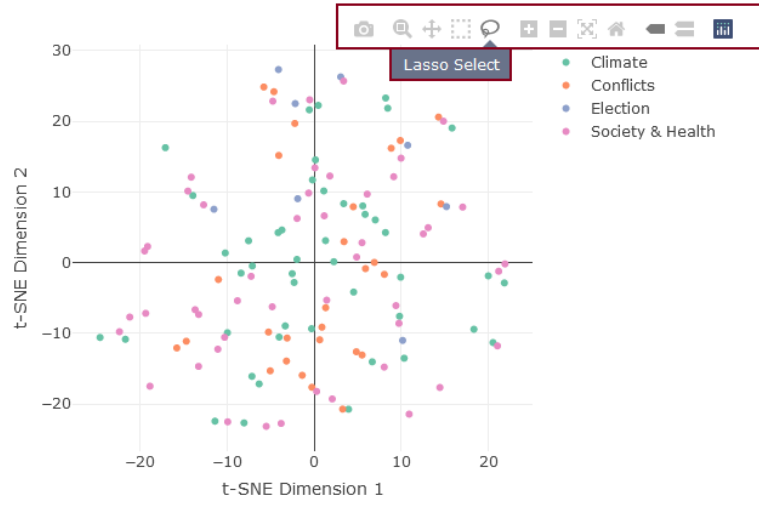


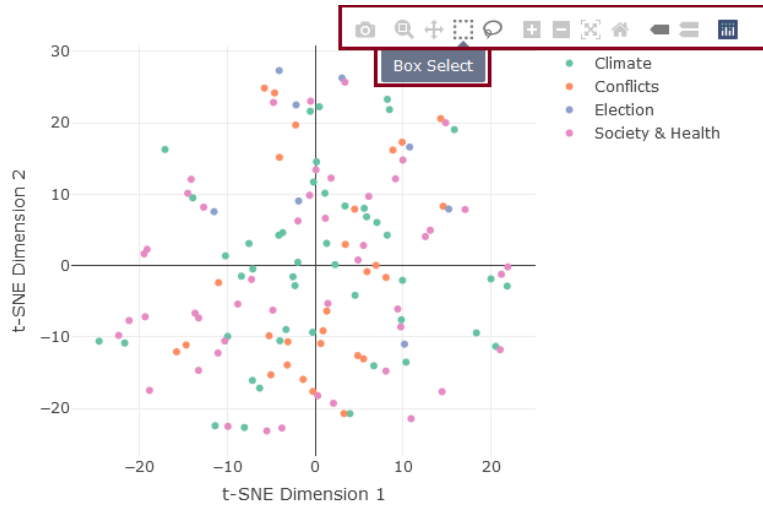
Figure 4.5: t-SNE Scatter Plot

In Figure 4.5 each dots represent single data point and distance between them represents degree of similarity between the data points (refer Section 2 for details). Users can interactively hover over the data points to see involved dimensions of data which is highlighted through tool-tip. The coloring of the cluster points is fixed and represents cluster membership to underlying dimension. By incorporating a color scheme to differentiate cluster points, the visual distinctiveness is enhanced, aiding users in identifying commonalities across various clusters HUND et al. (2016).

Subsequently, the user can choose to select either a single or multiple clusters using the mouse to brush over cluster points through either lasso selection or box select, enabling a deeper understanding of the cluster members and their respective subspaces, as shown by Figure 4.6a & 4.6b and elaborated further in next section.



(a) Lasso Select



(b) Box Select

Figure 4.6: t-SNE Scatter Plot Subspace selection

4.2.2 Filtering and re-computation

Users can select subset of cluster points and can investigate different properties of the selected subset. It provides information such as which dimen-

sion it belongs to, which enhances users understanding of data and to explore similar group of clusters as shown by Figure 4.7.

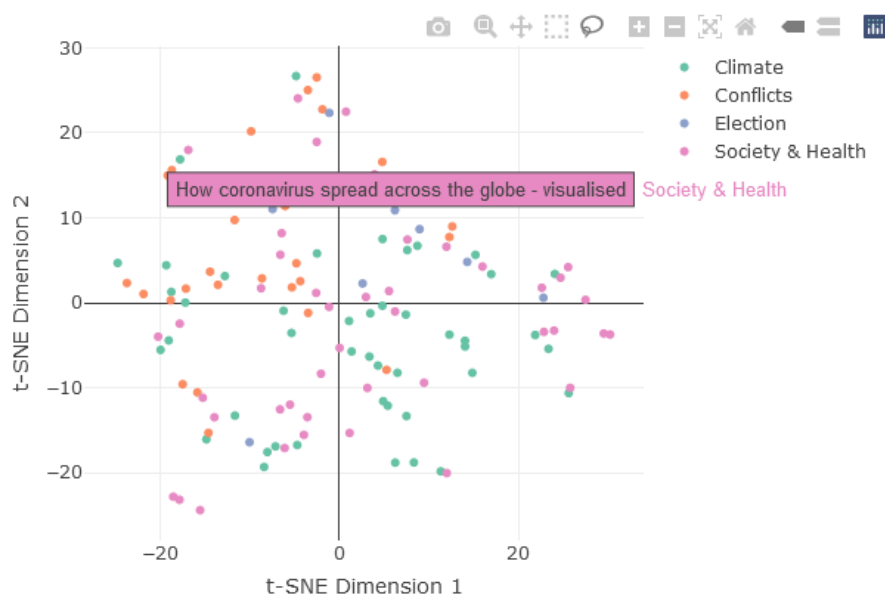


Figure 4.7: t-SNE Scatter Plot with additional dimensional information encoded by Hovering

Users can re-run the t-SNE algorithm with interactive adjustment parameters provided as sliders, namely *Number of Iterations* and *Perplexity* as shown by Figure 4.8. This enable users to explore more local subspace patterns from the data.

Apply Hierarchical Normalization:
☒ No
☐ Yes

Maximum number of iterations for the optimization

Number of iterations

5005,000

5009501,4001,8502,3002,7503,2003,6504,1004,5505,000

Perplexity controls the balance between attention to local and global aspects of the data. A higher perplexity value results in more emphasis on preserving the global structure of the data, while a lower perplexity value places more emphasis on preserving the local structure of the data.

Perplexity

530

59131721252930

Figure 4.8: t-SNE parameters to enhance and re-run algorithm

4.2.3 Deviation of Subspace plot

Upon selection of subset, users can further analyse by understanding statistical properties of subset. Figure 4.9 represents pairwise similarity of the selected subset and provides detailed understanding of how much the

subset deviates from the average. The provides insights to explore dimensional property of the cluster points.

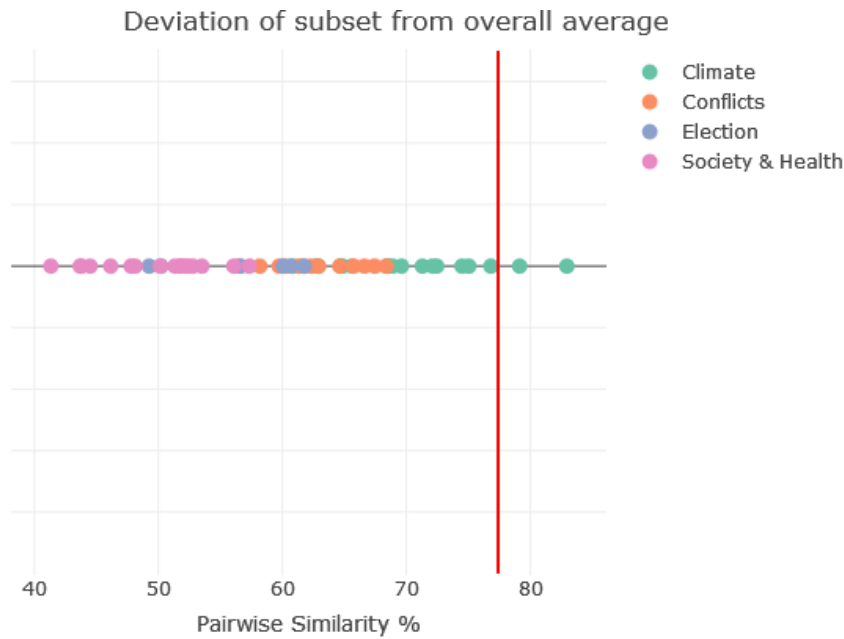


Figure 4.9: Plot represents deviation of the selected t-SNE subset from the average; uses pairwise similarity as calculation

4.2.4 Usage deviation plot

The Usage deviation plot as shown in Figure 4.10, represents usage in percentage of all the dimensions for given subest. Each column in Y-axis represents individual dimensions and the color coding represents the usage of single dimension in relation from the average. The dashed line of a dimension represents the average usage of the respective dimension in the whole dataset, whereas the solid line represents usage of the respective dimension by selected subset. The green color represents usage of dimensions above average, whereas red color represents usage of dimensions below average. This aids user in identifying relationships among all the dimensions for selected subset and helps identify underlying patterns within the selected subset.

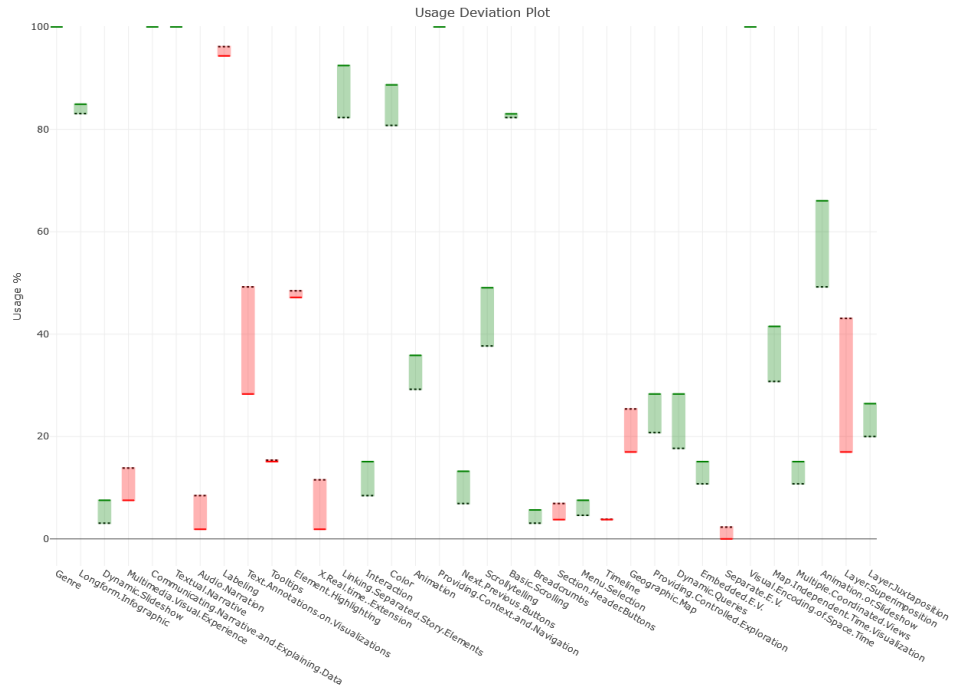


Figure 4.10: Plot represents

4.3 Interactive Visualizations: Bi-Clustering

4.3.1 General Overview: Bi-Clusters

Figure 4.11 provides an overview of the number of bi-clusters detected using the Bi-clustering Algorithm. In this scenario, each dot in the scatter plot represents a 2-dimensional representation of a bi-cluster. The position of the cluster points indicates the size of the corresponding bi-cluster, determined by its number of rows and columns. Bi-clusters are color-coded based on their quantity, and this color scheme remains

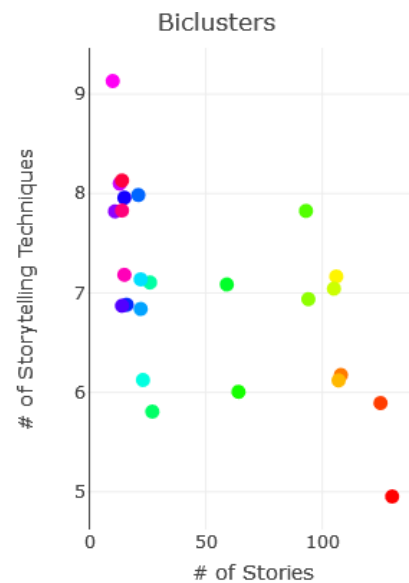
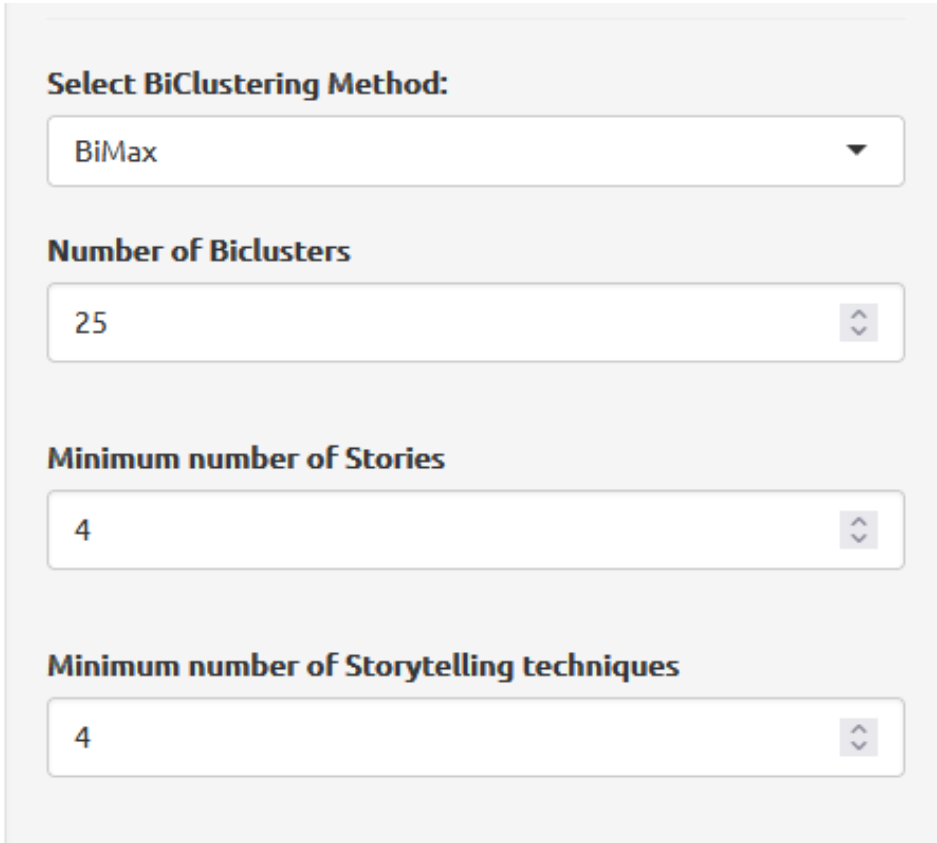


Figure 4.11: Projection of Biclusters using Scatter Plot

consistent across all three graph representations of bi-clustering discussed in subsequent sections.

4.3.2 Filtering and re-computation

Users can select number of bi-clusters and minimum number of rows and columns they intend as shown by Figure 4.12. This give users a degree of freedom to explore quanity and quality of patterns and interesting biclusters. These parameters then helps re-run the Bi-clustering algorithm until users find and coverage interseting pattern or dimensional relationship.



Select BiClustering Method:

BiMax

Number of Biclusters

25

Minimum number of Stories

4

Minimum number of Storytelling techniques

4

Figure 4.12: Parameters of Bi-clustering algortihm

4.3.3 Dendrogram Plot

The dendrogram shown in Figure 4.13 shows one of the possibilities to have an overview of biclusters. The dendrograms show all projection of biclusters calculated using agglomerative hierarchical clustering using Jaccardian Similarity Index. The node connection represents the Jaccardian Similarity between different biclusters. Each bicluster is assigned a color and this remains in sync with scatter plot and heatmap. Dissimilar to work of SEO und SHNEIDERMAN (2005); VENTOCILLA und RIVEIRO (2020) in Hierarchical Clustering Exploration (HCE), the dendrogram doesn't rank-by-feature whereas it represents similarity of biclusters and ranks them accordingly. This aids users with an apriori information when exploring the bicluster in detail.

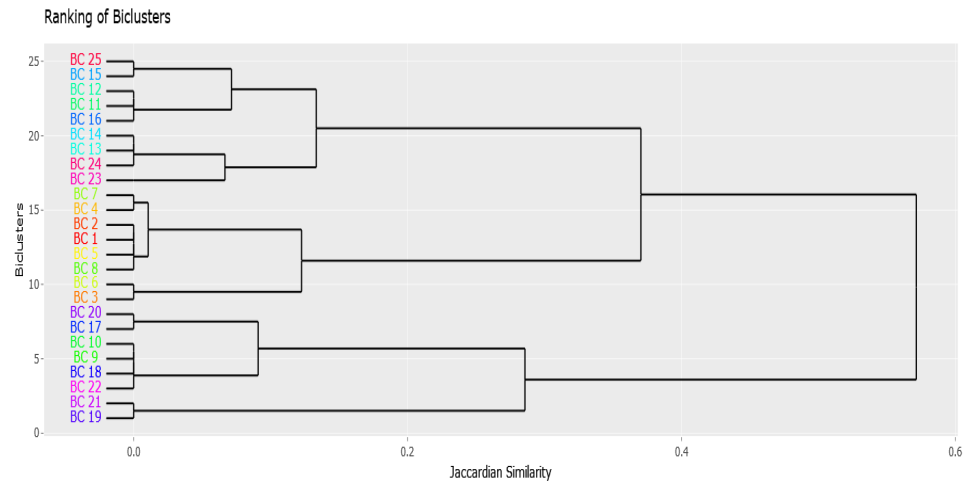


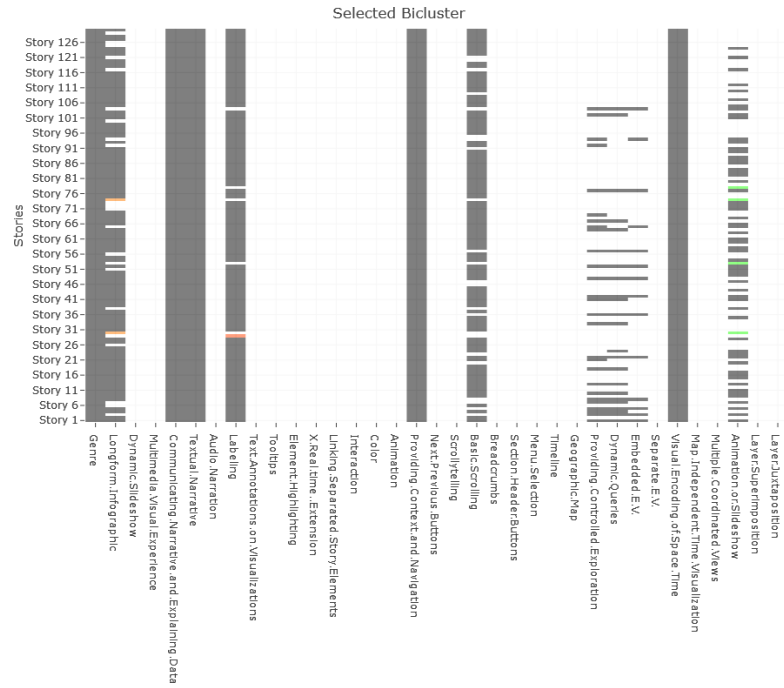
Figure 4.13: Ranking of Biclusters using color-encoded Dendrogram Plot

4.3.4 Heatmap

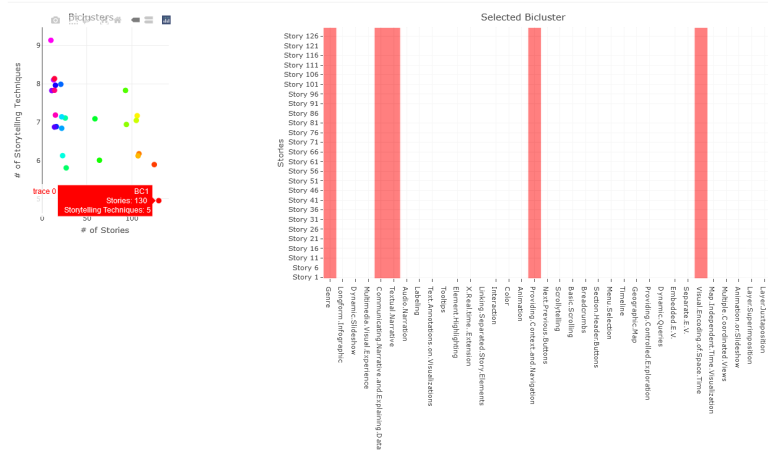
To represent details of selected bicluster, Figure 4.14a represents it through utilization of heatmap. Heatmaps offer an intuitive means of exploring the outcomes of clustering analysis by utilizing visual cues to showcase sample clustering and the variation in feature abundances across different cluster points ZHOU et al. (2021).

Each row and column represents 2-dimensional view of the raw-data. Each cell represents binary information whether the cell is part of the bi-

cluster or not. The colored cell of the heatmap represent the presence of cluster found within the data by biclustering algorithm. The gray color represent overlap of multiple biclusters as shown in Figure 4.14a. The color of the cell corresponds to the color of selected bicluster in scatter plot (details see Figure 4.11 and 4.14b), which indicates users section of the data, from which the biclusters are found. This positional information give a binary idea of a cluster found from the data and helps users to understand patterns and relationship of biclusters to the raw-data in detail.



(a) Projection of all overlapping Biclusters using Heatmap



(b) Projection of Biclusters using selection in Scatter Plot and its subsequent selection detail in Heatmap

Figure 4.14: Projection of Biclusters using Heatmap

5

Evaluation

The developed application is validated in an informal user study evaluation method with four domain experts from the Chair of Visualization at Otto-von-Guericke University. Evaluation also follows user-survey for usability evaluation. At last, section discusses analysis of results and promotes the ground for usability and the potential of the developed visual analytics application.

5.1 Evaluation Overview

Evaluation is one of the crucial and central part of visual analytics with importance equal to developing a visual analytics application itself . Evaluation is also key ingredient to turn visual analytics in to a mature technology.

5.2 Evaluation Set-up

As discussed in Section 4, a web-based visual analytics application has been developed. This application was evaluated on Storytelling Design-space curated by STEINHAEUER (2022) consisting of 130 Stories and 35 Storytelling techniques (refer Section 2 for details). The evaluation was performed by inviting four visual analytics domain experts from Chair of Visualization from Otto-von-Guericke University. For each of the four participants, a Zoom session was scheduled and recorded of 1 Hour/person. The sessions were planned to be evaluated in an informal setting following three step plan:

1. Demonstration of the dataset and application

2. Open Exploration

3. User-experience Survey

In Demonstration phase, participants were given a brief introduction of the dataset by introducing size and hierarchical complexity of the data. This was very brief phase with introduction taking upto 5-10 mins. Participants were introduced to the concepts of narrative visualization at a very brief to provide context of the dataset to be evaluated. After that, participants were given a demonstration of application on its use, provided graphs, and parameters support by the visual analytics application. Participants were also demonstrated details of linkage of graphs in the application.

In Exploration phase, participants explored the tool freely and independently. Participants were encouraged to think-out-loud while exploring, as the session was recorded for the analysis purpose. Participants were guided when in doubt by providing context of data or interactivity of the application. Participants explored the tool for 40-45 mins. They were not asked to perform specific set of tasks, rather based on the demonstration of dataset, they were asked to freely use the tool and explore and find if they can detect any patterns or relationship from the data using the application.

Later for the last 10 mins session, users were asked to fill user-experience survey. The survey was created with predefined set of questions. As suggested by KEIM et al. (2008a), the survey intended to capture three things in general:

1. perceived ease of use
2. perceived usefulness of the tool
3. user satisfaction

5.3 Evaluation Discussion

5.3.1 Analysis of User-studies

Four user-study sessions were scheduled with four study participants. For anonymization, they are named P1, P2, P3 and P4. This section discusses the user-experience and findings of storytelling data using the application by all four participants and tries to keep the observation of process and evaluation as objectively as possible.

Effectiveness of the application

Participants explored the storytelling dataset in Scatter Plot provided in the t-SNE tab. Participants were able to identify stories with the story title as text and theme of the story as color of the data point encoded in the view found upon hovering. Participants tried to look for clear clusters with respect to stories belonging to single theme but found it to be hard to find in the scatter plot. Therefore, the nature of the dataset didn't provide with natural clusters, however working with various parameters help find interesting clusters. One of the participants P2, selected a cluster to explore and found all the stories to be negative in nature and contained certain geographical aspect to it, although belonging to diverse Themes.

It was found that in t-SNE tab, the plot representing deviation of subset from the average didn't provide enough information on interestingness of the data-points or cluster in general. It was also found that due to linking and brushing property of the graphs, especially linking between scatter plot and usage deviation plot participants were able to get a detailed understanding of clusters and the relationships each story with storytelling technique.

With one of the exploration by P1, it was found that with multiple selection of cluster subsets, Visual Encoding of Space and Time and Interaction was used by stories a lot in combination Animation or Slideshow.

It is important to note that during exploration, participants found quite helpful and were able to detect the main categories and their usage in selected subset using usage deviation plot. The categories are *Genre*, *Communicating Narrative and Exploring Data*, *Linking Separate Story*

Element, Providing Context Navigation and Visual Encoding of Space and Time STEINHAUER (2022). Participants found that usage of all of these techniques were nearly 100%, and they were found to be important when finding patterns. This could also be the case partly due to influence of *apriori* information about the hierarchy of the storytelling techniques.

An interesting thing found by P1 was when a subset was selected with no stories that belonged to Conflict, that gave a better understanding of Conflict stories. It made it easy to understand that non-Conflict stories subset used very less Multimedia Visual Experience and Audio Narration (0%). Therefore an interesting pattern was found and P1 interpreted as

"when stories with Conflict were are written, they are very straightforward and it makes more sense that less freedom to play around with the data is given and stories do not support more subjective interpretation due to sensitivity of the topic".

Participant P2 found the selection of subset in the scatter plot to be interesting as selected cluster contained stories from topic of Risk although belonging to different Themes.

"It is interesting to see that a random selection of cluster showed me topics concerning to topics around Risk. Although all the data points tells me diverse types and categories of Risk".

P2 also found that pairwise similarity of the subset was quite similar to the overall average of all the data points.

Additionally P3, also found that in general by selection of large clusters, Conflict data was found to be much similar to the average compared to other Themes. P4 also confirmed that Election stories were quite dissimilar to average. This gave an overall idea of scatter plot representing subset similarity from overall average to be less significant in looking for direct patterns.

Some participants also found it quite difficult to find patterns using t-SNE tab, since it required more options such as filtering and sorting options for Themes to understand underlying relationship. P3 also mentioned,

"with t-SNE it is tough to provide any information on whether different Themes have common techniques?"

All participants found the use of Hierarchical Normalization as very useful parameter to find patterns. This added a layer of clarification of clusters upon the t-SNE projections. With normalization, participants were able to visualize clearer clusters and were easily able to identify patterns. For example, P1 was able to find clear usage of *Multimedia Visual Experience* technique along with *Animation* and *Color* in a subset which was tough to identify before normalization. P4 also mentioned,

"It is clear to see clusters by applying hierarchical normalization, although clusters are from diverse Themes, which speaks good of the algorithm approach".

Normalization also gives denser cluster with low perplexity and scattered cluster with high perplexity. According to P1, Hierarchical Normalization provides distinguished groups of clusters, which in turn builds trust in the data as well as cluster associations. Therefore, the normalization of hierarchies aid clearer cluster and more interesting patterns from the dataset.

For the second section of analysis, participants evaluated patterns using the Bi-Clustering tab. In general, participants found less clearer patterns to be evaluated. Participants found interesting to see ranking of found bi-clusters using dendrogram plot but found less helpful because of missing linkage of dendrogram plot to scatter plot and heatmap.

Participants found found biclusters interesting and tried exploring individual bi-clusters. One of the participants found Providing Controlled Exploration being under-reported in Bi-clustering Heatmap compared to patterns found by t-SNE. Because participants used t-SNE tab first to detect patterns and then used Bi-Clustering tab, they found it being two different analysis rather than one analysis with two approaches. Participants found it hard to comprehend clusters found by biclusters in a heatmap. One of the reason mentioned was the binary aspect of the algorithm, showing either a cluster is present or it isn't. Therefore, it generates a need for a more fuzzier logic of the bi-clustering algorithm to be used.

Due to overview first and details on demand principle, participants found it straightforward to see all the found bi-clusters in a scatter plot and details of bicluster generated in Heatmap by selecting individual cluster point. However, more encoding of information was expected missing. Par-

ticipants found it hard and found that if story name and Themes encoding would make it very helpful to find patterns.

P3 believed, *"at the moment it is still hard to find patterns with t-SNE due to complexity similar to Multi-dimensional Scaling and hope this complexity is resolved by using Bi-clustering but with additional extension of information encoding"*

Analysis of user-satisfaction study

Analysis of user-satisfaction was done using survey forms presented after each Zoom session as mentioned in Section 5.2. Survey forms followed a set of questionnaire divided into 5 sections and all the section follow LIKERT (1932) for to answer a question except for few descriptive questions as mentioned below:

1. User's prior knowledge of the application and it's integrants
2. User's Experience with tool
3. Algorithm specific questions
4. Pattern-detection questions
5. Additional comments

To evaluate User's experience with the tool, SUS Score was evaluated JORDAN et al. (1996). The average SUS score was found to be **77.5**. According to the grading represented by Table 5.1 below, the application received a *Good* Score with Grade *B*. This represents that user's were quite satisfied with the application but it still requires further improvements and validation.

SUS Score	Grade	Adjective Rating
>80.3	A	Excellent
68-80.3	B	Good
68	C	Okay
51-68	D	Poor
<51	F	Awful

Table 5.1: SUS Score grading

5.4 Limitations

The current state of application allows user to detect patterns from hierarchical data with shallow hierarchy similar to storytelling dataset using t-SNE and Bi-clustering algorithm based views. However, following are the discussed known limitation:

- **Filtering and Sorting of Dimensions:** In order to explore found patterns using t-SNE, Usage Deviation Plot does not provide any option to filter or sort based on Themes.
- **Comparison of Biclusters:** As described in Section 4, scatter plot represents set of all biclusters found by algorithm. However, more than one biclusters cannot be selected as each biclusters have individual color assigned from color scale and selecting more than one would show the overlapping Biclusters as just grey area in heatmap.
- **Bidirectional filtering** Currently application only allows users to applying linking of graphs based on "Overview first and Details on demand" SHNEIDERMAN (1996). However using this principle, the subset selection and linking is only applied unidirectional, i.e., Overview to details.
- **Binary nature of Biclustering:** In order to find pattern using biclustering algorithm, the graphs represent a binary aspect of the clusters in Heatmap, i.e., whether the cluster is present in the data or not and this is by limited nature of Bi-clustering algorithm.

6

Conclusions and Future Work

6.1 Conclusions

Pattern detection is a challenging task especially in hierarchical data with shallow hierarchy. Therefore, this work proposes a web-based visual analytics application that can be used to detect patterns in such form of data. The application uses t-SNE and Bi-clustering algorithms to detect patterns. The application is evaluated on storytelling design space containing 130 stories and 35 storytelling techniques with one level hierarchical depth categorised into 6 main categories. The application was evaluated by four domain experts and showed that such an application can help analyst find patterns from the stories. It also suggested that application can help find new storytelling techniques when crafting new stories or developers identifying techniques that can be used when narrating stories and to develop tools around digital journalism. The application can be used for general datasets that follow similar structure of storytelling design space.

6.2 Future Work

In Section 5.4, the thesis addresses the limitations encountered throughout the study and identifies potential gaps that warrant further investigation. This work can be expanded by embarking on future research endeavors that begin with:

- **Investigate suitable default parameters for algorithms** One possibility could be investigate further for default of parameters used by t-SNE and Biclustering, as optimal parameters can converge to find patterns faster EREN et al. (2013); WANG et al. (2016).

- **Implement similarity based color coding** For biclustering algorithms, color coding can be implemented based on similarity of biclusters and hierarchical structure of data to enhance bicluster visualization TENNEKES und DE JONGE (2014); ZHOU und HANSEN (2016).
- **Investigate use of flexible biclustering algorithms** As mentioned in the section above, a flexible and non-binary restrictive algorithms such as BOZDAĞ et al. (2009); GU und LIU (2008); HOCHREITER et al. (2010); LI et al. (2009) can be examined to analyse more local patterns of the data.
- **Add story title information while finding similarity** As discussed in section above, textual information from story titles can be embedded into the similarity based algorithms or in a hybrid approach, which adds further layer of information to algorithm while calculating similarity of data points and creates more localised pattern from the data.
- **Hierarchical clustering for hierarchies** One possibility could be to investigate hierarchical clustering algorithm to investigate for data with shallow hierarchies to explore similarities caused by dependencies of dimensions.



Abbreviations and Notations

Dataset and clustering acronyms

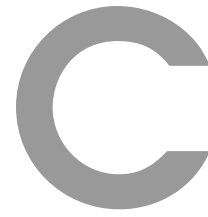
Acronym	Meaning
VA	Visual Analytics
t-SNE	t-Distributed Stochastic Neighbor Embedding
MDS	Multi Dimensional Scaling
PCA	Principal Component Analysis
Bimax	Binary inclusion-maximal biclustering algorithm

B

List of Figures

2.1	Storytelling design space analysis data curated by STEINHAUER (2022); 130 Stories and 35 Storytelling techniques . . .	7
2.2	Defining Visual Analytics , combination of different disciplines COOK und THOMAS (2005)	8
2.3	Visual Analytics process KEIM et al. (2010, 2008c)	9
2.4	Non-Linear Dimensionality Reduction in Visual Analytics process (inspired by PREIM (2023))	10
4.1	Visual Analytics Process Architecture	18
4.2	R-Shiny WebApp depicting different sections of the app. A) represents the drop-down menu to select a dataset. B) represents the control parameters of t-SNE algorithm; namely hierarchical normalization, number of iterations, and Perplexity as discussed in Section 2, C) represents tab to explore the algorithm, D) Scatter-plot for 2D representation of t-SNE projections, E) Scatter-plot with red line marker; the red line marker represents average pairwise similarity of the subset of data in plot D and the scatter markers represent the selected pairwise similarity of subset	19
4.3	R-Shiny Web-App Bi-Clustering	20
4.4	R-Shiny WebApp NANTASENAMAT (2020)	22
4.5	t-SNE Scatter Plot	22
4.6	t-SNE Scatter Plot Subspace selection	24

4.7 t-SNE Scatter Plot with additional dimensional information encoded by Hovering	25
4.8 t-SNE parameters to enhance and re-run algorithm	26
4.9 Plot represents deviation of the selected t-SNE subset from the average; uses pairwise similarity as calculation	27
4.10 Plot represents	28
4.11 Projection of Biclusters using Scatter Plot	28
4.12 Parameters of Bi-clustering algortihm	29
4.13 Ranking of Biclusters using color-encoded Dendogram Plot	30
4.14 Projection of Biclusters using Heatmap	32



List of Tables

5.1	SUS Score grading	38
-----	-----------------------------	----



Bibliography

- [BARNES und HUT 1986] J. Barnes und P. Hut. **A hierarchical $O(N \log N)$ force-calculation algorithm.** nature, Vol. 324(6096):446–449, 1986.
- [BECKER und CLEVELAND 1987] R. A. Becker und W. S. Cleveland. **Brushing scatterplots.** Technometrics, Vol. 29(2):127–142, 1987.
- [BOZDAĞ et al. 2009] D. Bozdağ, J. D. Parvin und U. V. Catalyurek. **A biclustering method to discover co-regulated genes using diverse gene expression datasets.** In: Bioinformatics and Computational Biology: First International Conference, BICoB 2009, New Orleans, LA, USA, April 8-10, 2009. Proceedings, pp. 151–163. 2009, Springer.
- [BUJA et al. 1991] A. Buja, J. A. McDonald, J. Michalak und W. Stuetzle. **Interactive data visualization using focusing and linking.** In: Proceedings of the 2nd conference on Visualization'91, 1991, pp. 156–163.
- [CHEN und CAFARELLA] **Automatic web spreadsheet data extraction.**
- [COOK und THOMAS 2005] K. A. Cook und J. J. Thomas. **Illuminating the path: The research and development agenda for visual analytics.** Tech. rep., Pacific Northwest National Lab.(PNNL), Richland, WA (United States), 2005.
- [CRAWLEY 2012] M. J. Crawley. **The R book.** John Wiley & Sons, 2012.
- [DEVASSY und GEORGE 2020] B. M. Devassy und S. George. **Dimensionality reduction and visualisation of hyperspectral ink data using t-SNE.** Forensic science international, Vol. 311:110194, 2020.

- [DOU et al.] **Expandable group identification in spreadsheets.**
- [ECKELT et al.] **TourDino: A Support View for Confirming Patterns in Tabular Data.**
- [EREN et al. 2013] K. Eren, M. Deveci, O. Küçüktunç und Ü. V. Çatalyürek. **A comparative analysis of biclustering algorithms for gene expression data.** Briefings in bioinformatics, Vol. 14(3):279–292, 2013.
- [FURMANOVA et al. 2017] K. Furmanova, S. Gratzl, H. Stitz, T. Zichner, M. Jaresova, M. Ennemoser, A. Lex und M. Streit. **Taggle: Scalable visualization of tabular data through aggregation.** arXiv preprint arXiv:1712.05944, Vol. 6, 2017.
- [GIORGI et al. 2022] F. M. Giorgi, C. Ceraolo und D. Mercatelli. **The R language: an engine for bioinformatics and data science.** Life, Vol. 12(5):648, 2022.
- [GREENE et al. 2014] C. S. Greene, J. Tan, M. Ung, J. H. Moore und C. Cheng. **Big data bioinformatics.** Journal of cellular physiology, Vol. 229(12):1896–1900, 2014.
- [GU und LIU 2008] J. Gu und J. S. Liu. **Bayesian biclustering of gene expression data.** BMC genomics, Vol. 9(1):1–10, 2008.
- [HINTON und ROWEIS 2002] G. E. Hinton und S. Roweis. **Stochastic neighbor embedding.** Advances in neural information processing systems, Vol. 15, 2002.
- [HOCHREITER et al. 2010] S. Hochreiter, U. Bodenhofer, M. Heusel, A. Mayr, A. Mitterecker, A. Kasim, T. Khamiakova, S. Van Sanden, D. Lin, W. Talloen et al. **FABIA: factor analysis for bicluster acquisition.** Bioinformatics, Vol. 26(12):1520–1527, 2010.
- [HUND et al. 2016] M. Hund, D. Böhm, W. Sturm, M. Sedlmair, T. Schreck, T. Ullrich, D. A. Keim, L. Majnaric und A. Holzinger. **Visual analytics for concept exploration in subspaces of patient groups.** Brain Informatics, Vol. 3(4):233–247, 2016.
- [JIA et al. 2022] L. Jia, W. Yao, Y. Jiang, Y. Li, Z. Wang, H. Li, F. Huang, J. Li, T. Chen und H. Zhang. **Development of interactive biolog-**

-
- ical web applications with R/Shiny.** Briefings in Bioinformatics, Vol. 23(1):bbab415, 2022.
- [JORDAN et al. 1996] P. W. Jordan, B. Thomas, I. L. McClelland und B. Weerdmeester. **Usability evaluation in industry.** CRC Press, 1996.
- [KEIM et al. 2008a] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer und G. Melançon. **Visual analytics: Definition, process, and challenges.** In: Information visualization, pp. 154–175. 2008. Springer.
- [KEIM et al. 2010] D. Keim, J. Kohlhammer, G. Ellis und F. Mansmann. **Mastering the information age solving problems with visual analytics.** Eurographics Association, 2010.
- [KEIM 2002] D. A. Keim. **Information visualization and visual data mining.** IEEE transactions on Visualization and Computer Graphics, Vol. 8(1):1–8, 2002.
- [KEIM et al.] **Visual analytics: Combining automated discovery with interactive visualizations.**
- [KEIM et al. 2008] D. A. Keim, F. Mansmann, J. Schneidewind, J. Thomas und H. Ziegler. **Visual analytics: Scope and challenges.** Springer, 2008c.
- [LI et al.] **HiTailor: Interactive Transformation and Visualization for Hierarchical Tabular Data.**
- [LI et al. 2009] G. Li, Q. Ma, H. Tang, A. H. Paterson und Y. Xu. **QUBIC: a qualitative biclustering algorithm for analyses of gene expression data.** Nucleic acids research, Vol. 37(15):e101–e101, 2009.
- [LIKERT 1932] R. Likert. **A technique for the measurement of attitudes.** Archives of psychology, 1932.
- [MERČUN et al.] **FrbrVis: An information visualization approach to presenting FRBR work families.**
- [NAM et al.] **Clustersculptor: A visual analytics tool for high-dimensional data.**
- [NANTASENAMAT] **Build Your First Shiny Web App in R.**

- [PERIN et al. 2014] C. Perin, P. Dragicevic und J.-D. Fekete. **Revisiting bertin matrices: New interactions for crafting tabular visualizations**. IEEE transactions on visualization and computer graphics, Vol. 20(12):2082–2091, 2014.
- [PILLAT et al.] **Experimental study on evaluation of multidimensional information visualization techniques**.
- [PREIM] **Non Linear Dimensionality Reduction, Visual Analytics Lecture**.
- [PRELIĆ et al. 2006] A. Prelić, S. Bleuler, P. Zimmermann, A. Wille, P. Bühlmann, W. Gruissem, L. Hennig, L. Thiele und E. Zitzler. **A systematic comparison and evaluation of biclustering methods for gene expression data**. Bioinformatics, Vol. 22(9):1122–1129, 2006.
- [SAAKE et al. 2000] G. Saake, K.-U. Sattler und D. A. Keim. **Datenbank-und Visualisierungstechnologien in der Informationsfusion**. In: SimVis, 2000, pp. 1–14.
- [SALVANESCHI et al.] **Reactive programming: A walkthrough**.
- [SCHMULLER 2017] J. Schmuller. **Statistical Analysis with R For Dummies**. John Wiley & Sons, 2017.
- [SEGEL und HEER 2010] E. Segel und J. Heer. **Narrative Visualization: Telling Stories with Data**. IEEE Transactions on Visualization and Computer Graphics, Vol. 16(6):1139–1148, 2010.
- [SEO und SHNEIDERMAN 2005] J. Seo und B. Shneiderman. **Knowledge discovery in high dimensional data: case studies and a user survey for an information visualization tool**. Tech. rep., 2005.
- [SHNEIDERMAN 1996] B. Shneiderman. **The eyes have it: a task by data type taxonomy for information visualizations**. In: Proceedings 1996 IEEE Symposium on Visual Languages, 1996, pp. 336–343.
- [STEED et al. 2020] C. A. Steed, J. R. Goodall, J. Chae und A. Trofimov. **CrossVis: A visual analytics system for exploring heterogeneous multivariate data with applications to materials and climate sciences**. Graphics and Visual Computing, Vol. 3:200013, 2020.

-
- [STEINHAEUER 2022] N. Steinhauer. **A Descriptive Characterization of Interactive Data-driven Visual Storytelling in a Spatio-temporal Context**. Master's thesis, Dept. of Computer Science, 2022.
- [STOLPER et al.] **Emerging and recurring data-driven storytelling techniques: Analysis of a curated collection of recent stories**.
- [STREIT et al. 2019] M. Streit, S. Gratzl, H. Stitz, A. Wernitznig, T. Zichner und C. Haslinger. **Ordino: a visual cancer analysis tool for ranking and exploring genes, cell lines and tissue samples**. *Bioinformatics*, Vol. 35(17):3140–3142, 2019.
- [TENNEKES und DE JONGE 2014] M. Tennekes und E. de Jonge. **Tree Colors: Color Schemes for Tree-Structured Data**. *IEEE Transactions on Visualization and Computer Graphics*, Vol. 20(12):2072–2081, 2014.
- [VAN DER MAATEN 2014] L. Van Der Maaten. **Accelerating t-SNE using tree-based algorithms**. *The journal of machine learning research*, Vol. 15(1):3221–3245, 2014.
- [VAN DER MAATEN und HINTON 2008] L. Van der Maaten und G. Hinton. **Visualizing data using t-SNE**. *Journal of machine learning research*, Vol. 9(11), 2008.
- [VENTOCILLA und RIVEIRO 2020] E. Ventocilla und M. Riveiro. **A comparative user study of visualization techniques for cluster analysis of multidimensional data sets**. *Information visualization*, Vol. 19(4):318–338, 2020.
- [VOIGT 2002] R. Voigt. **An extended scatterplot matrix and case studies in information visualization**. Master's thesis, Hochschule Magdeburg-Stendal, 2002.
- [WAGEMANS et al. 2012] J. Wagemans, J. H. Elder, M. Kubovy, S. E. Palmer, M. A. Peterson, M. Singh und R. Von der Heydt. **A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure-ground organization**. *Psychological bulletin*, Vol. 138(6):1172, 2012.
- [WANG et al. 2016] Z. Wang, G. Li, R. W. Robinson und X. Huang. **UniBic: Sequential row-based biclustering algorithm for analysis of gene expression data**. *Scientific reports*, Vol. 6(1):1–10, 2016.

- [WICKHAM et al. 2023] H. Wickham, M. Çetinkaya-Rundel und G. Grolemund. **R for data science**. " O'Reilly Media, Inc.", 2023.
- [WONG und THOMAS 2004] P. C. Wong und J. Thomas. **Visual analytics**. IEEE Computer Graphics and applications, Vol. 24(05):20–21, 2004.
- [XU et al. 2020] X. Xu, Z. Xie, Z. Yang, D. Li und X. Xu. **A t-SNE Based Classification Approach to Compositional Microbiome Data**. Frontiers in Genetics, Vol. 11, 2020.
- [YIANILOS 1993] P. N. Yianilos. **Data structures and algorithms for nearest neighbor**. In: Proceedings of the ACM-SIAM Symposium on Discrete algorithms, Vol. 66, 1993, p. 311.
- [ZHANG et al.] **A visual analytics approach for the diagnosis of heterogeneous and multidimensional machine maintenance data**.
- [ZHOU et al. 2021] G. Zhou, J. Ewald und J. Xia. **OmicsAnalyst: a comprehensive web-based platform for visual analytics of multi-omics data**. Nucleic Acids Research, Vol. 49(W1):W476–W482, 2021.
- [ZHOU und HANSEN 2016] L. Zhou und C. D. Hansen. **A Survey of Colormaps in Visualization**. IEEE Transactions on Visualization and Computer Graphics, Vol. 22(8):2051–2069, 2016.
- [ZHOU et al. 2017] Z. Zhou, Z. Ye, Y. Liu, F. Liu, Y. Tao und W. Su. **Visual analytics for spatial clusters of air-quality data**. IEEE computer graphics and applications, Vol. 37(5):98–105, 2017.

Declaration of Academic Integrity

I hereby declare that I have written the present work myself and did not use any sources or tools other than the ones indicated.

Date: 19.06.2023

.....

(Signature)