Unsupervised and Dynamic Dendrogram-based Visualization of Medical Data

Angela Moufarrej[1], Abdulkader Fatouh[1], and Joe Tekli[1] ✉

[1] E.C.E. Department, Lebanese American University, 36 Byblos, Lebanon
{angela.moufarrej, abdulkader.fatouh}@lau.edu, joe.tekli@lau.edu.lb

Abstract. Visualizing the correlations between structured medical data in the form of Electronic health records (EHRs) is of major importance for effective and efficient medical data analysis and decision-making. This work describes an unsupervised semi-structured and feature-based tool for dynamic EHR data visualization called "mirrored dendrograms". It accepts as input semi-structured EHRs, and allows the user to select the target features to be visualized and mapped against each other, and their relative weights on the visualization process. It then invokes a hierarchical clustering process to cluster the data following the user-chosen features, and produces a dendrogram structure for each combination of target features. The dendrograms are mirrored against each other by mapping their nodes using the transportation optimization problem, allowing the user to dynamically zoom-in and out of the mapping at different granularity levels. We have evaluated our solution using a sample dataset of 114 EHRs of patients who suffer from migraine disorder. A group of 20 testers participated in the evaluations to assess the tool compared with existing solutions. Results showcase the tool's performance.

Keywords: Data Visualization, Data Clustering, Dendrogram, Feature Correlation, Similarity Computation, Data Granularity.

1 Introduction

Extracting and understanding the correlations between data features is increasingly important in many applications, ranging over business, demographics, politics, and more specifically medicine [5, 7, 19]. The proper exploitation of medical data introduces many challenges in terms of data analysis and visualization, to allow effective and efficient decision-making. The problem is further aggravated on the Web where medical data is often loosely structured and multi-featured. In this context, interactive data visualization comes into play as a promising solution to facilitate data analysis. Data visualization allows unveiling patterns and trends that could be repeated over time and space, and helps experts identify anomalies in the data [19, 23].

This work describes a new unsupervised tool for dynamic data visualization called *mirrored dendrograms*. It accepts as input semi-structured medical data in the form of Electronic Health Records (EHRs) and allows the user to select the target features to be visualized and mapped against each other. A hierarchical clustering process is invoked to cluster the data and produce a dendrogram structure for each combination of target features. The tool recommends the best zooming level to display the dendrograms, highlighting the maximum correlation (similarity) with the minimal amount of details (granularity) presented to the user. This is based on our intuition that users wish to acquire the most value out of the data while viewing the least amount of data, i.e., with the least amount of effort. The dendrograms are then mirrored against each other, where

their leaf nodes and inner nodes are mapped against each other, identifying the best connections using the transportation optimization problem. The user can dynamically adjust the zooming level to zoom-in and out of the mapping at different granularity levels. Different from existing solutions like tanglegrams and heatmap dendrograms, our work offers three main contributions: (i) connecting the dendrograms through their internal nodes to describe their structure relationships (instead of connecting their leaf nodes only), (ii) allowing to zoom-in and out the data to show their relationships at different granularity levels (compared with existing static solutions), and (iii) identifying the best zooming level between the two dendrograms which highlights the maximum correlation with the minimal amount of details presented to the user (acquiring the most value out of the data, while viewing the least amount of data).

We have evaluated our solution using a sample dataset of 114 EHRs of patients who suffer from migraine disorder. A group of 20 testers participated in the evaluations to assess the tool compared with existing solutions. Results showcase the tool's performance.

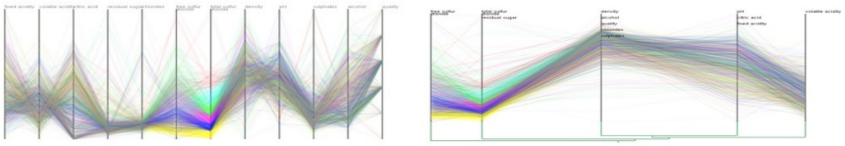
The remainder of this paper is organized as follows. Section 2 reviews related visualization tools. Section 3 describes the proposal. Section 4 presents the experimental evaluation and results, before concluding in Section 5 with future works.

2 Related Work

We provide a brief review of visualization tools based on clustering techniques, including parallel coordinates, dendrogram, tanglegram, and heatmap visualizations.

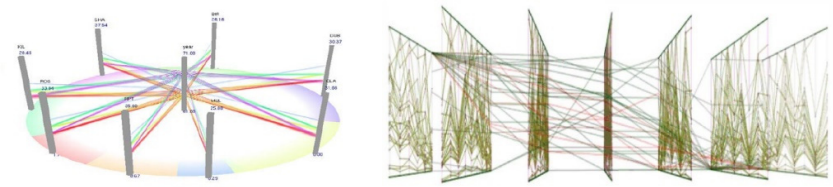
2.1. Parallel Coordinates

Parallel coordinates is a common visualization technique that aims at representing multi-dimensional datasets and extracting the underlying relationships between them (cf. Figure 1). In an *N*-dimensional space, a single data element is plotted as a polyline that crosses the *N* vertical axes, where its location on each axis is proportional to its value for the dimension related to that axis. Data points on adjacent axes are linked together, highlighting the correlation between the dimensions. While effective with relatively small datasets, yet this technique can suffer from cluttering when dealing with large data samples and dimensions [4, 18]. The authors in [18] propose a solution based on the concept of contractible parallel coordinates, suggesting to merge highly correlated vertical axes together (cf. Figure 1.b). This requires reordering the vertical axes to get the most correlated ones next to each other, by computing pair-wise correlations between all data dimensions, and then merging the most correlated ones together into a single vertical axis.



a. Original parallel coordinates representation **b.** Reduced representation following [18]

Figure 1. Sample parallel coordinates representations



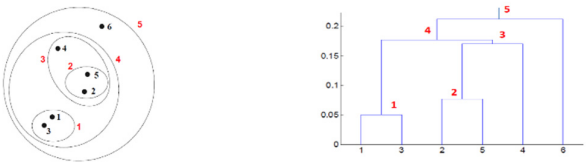
a. Sample visualization by [16] **b.** Sample visualization by [31]

Figure 2. Sample 3D visualizations based on parallel coordinates tool

In [16], the authors extend the usual 2D display of parallel coordinates and introduce a new 3D visualization tool that allows visualizing the correlations between several features at a time (cf. Figure 3.a)., compared with the traditional 2D display which can only visualize the correlation between two dimensions at once. It enables analyzing concurrently one-to-one relations between a central "focus" dimension and the remaining dimensions situated around it, forming a cylinder. In [31], the authors extend the parallel coordinates tool to add a 3D visualization considering the time dimension (cf. Figure 2.b). They include multiple planes each representing a certain time stamp. This forms a group of plane clusters where each plane includes the parallel coordinates visualization depending on the timestamp of the data samples, where data sampled at the same time is represented on the same plane.

2.2. Dendrogram

A dendrogram is a diagram representing a tree that shows the hierarchical relationships between data points or objects [2]. It consists of a hierarchy of clusters where the leaf nodes represent individual data points, the internal nodes represent clusters of data points, and the root node represents the entire data set (cf. Figure 3). Dendrograms provide a visual description, i.e., an explanation of the hierarchical clustering process and how the clusters were formed, compared with other clustering techniques like partitional clustering or spectral clustering where no such explanation or visualization exists to describe the clustering process [27, 28].



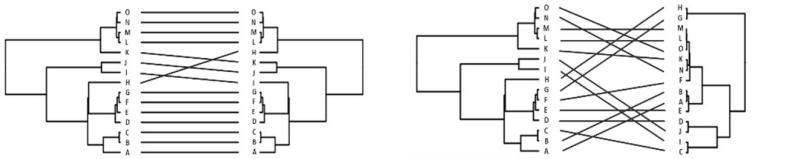
a. Hierarchical clustering b. Dendrogram structure describing the clusters in (a)

Figure 3. Sample hierarchical clustering and corresponding dendrogram structure

2.3. Tanglegram

A tanglegram allows comparing two pairs of dendrograms (cf. Figure 4). The aim is to reduce the number of line crossings (known as entanglements) to make the visualization clearer and easier to understand [6, 8]. Fewer (higher) crossings between the tree leaves might indicate higher (lower) correlation between the tree structures. Yet the trees being compared can have different internal structures or topologies, while their leaf nodes are

presented in a matching order. This can be misleading when evaluating correlation between tree structures [8].

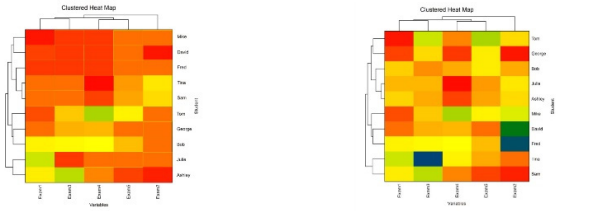


a. Few crossings (high correlation between leaves) b. Many crossings (low correlation between leaves)

Figure 4. Sample tanglegram representations based on [8]

2.4. Cluster Heatmap

Cluster heatmap shows two dendrograms in a data matrix, one positioned as row and the other one positioned as column (cf. Figure 5). Rows and columns may be perceived to be highly or poorly correlated according to the ordering of their dendrogram leaf nodes, which can be misleading [20] (similarly to tanglegrams). Also, when clusters are formed close to the root of the dendrogram, cells that are not closely clustered must still be placed adjacent in the heatmap due to the rigid grid structure [14]. Hence, rows or columns that are closely clustered can also end up non-adjacent in large clusters [13].



a. Cluster heatmap showing highly correlated rows and columns

b. Cluster heatmap showing low correlation between row and columns

Figure 5. Sample cluster heatmap visualizations from [17]

Few alternatives have been suggested to compensate for the limitations of cluster heatmaps [13], including circle packing, sunburst, and radial dendrogram (cf. Figure 6). Yet most of them aim at improving the visualization of the clusters within an individual dataset, and do not allow comparing pairs of datasets.

a. Circle packing [29] b. Sunburst [24] c. Radial dendrogram [13] d. Force directed tree [9]

Figure 6. Suggested alternatives to cluster heatmaps

3. Proposal

We design a new tool for interactive visualization of structured data titled *mirrored dendrograms* (cf. overall architecture in Figure 7). It accepts as input two EHRs represented as sets of semi-structured and multi-featured data, and allows the user to select the target features to be visualized. The data is then hierarchically clustered to produce a dendrogram for each combination of target features. The tool evaluates the structural similarity between the produced dendrograms to identify the best zooming level to display the data. The dendrograms' internal nodes are mapped against each other using an adaptation of the transportation optimization problem. The tool allowing the users to dynamically adjust the zooming level, and the number and weight of the connections, according to their needs.

3.1. Data Representation

We use real-world EHRs to describe our running examples, yet any other multi-featured data can be utilized. Figure 8 shows extracts of two EHRs providing atomic feature elements (e.g., *DOB*, *days of migraine*, *age at onset*) and aggregate feature elements (e.g., *personal information*, *migraine data*, *vital signs*).

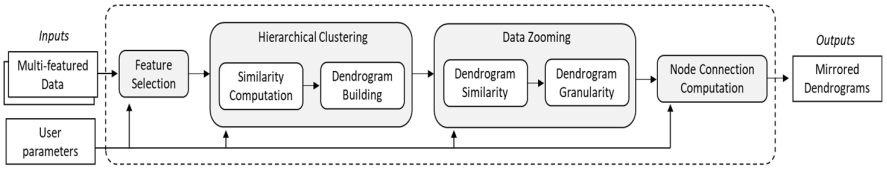
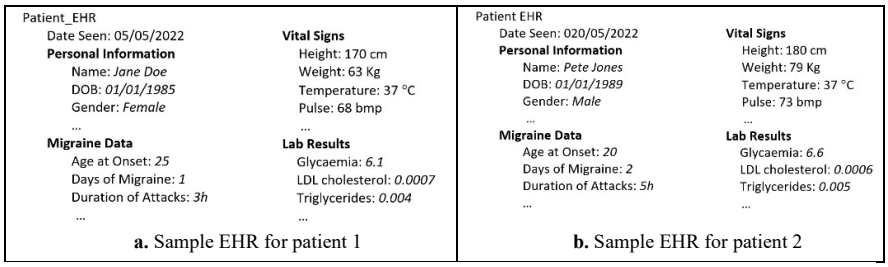


Figure 7. Simplified activity diagram describing our approach's overall architecture



a. Sample EHR for patient 1 **b.** Sample EHR for patient 2

Figure 8. Sample EHRs for two migraine patients

3.2. Similarity Computation

After identifying the features of interest, the next step is to perform feature similarity computation to conduct hierarchical clustering. Similarity between atomic feature elements are computed according to their feature data-types (Table 1). Similarity between aggregate feature elements is computed as the aggregation of the similarities of their constituent atomic elements. This can be computed in several ways, using for instance the *maximum*, *minimum*, *average*, or *weighted sum* functions [26, 28]. Here, we make use of the *weighted sum* function since it enables the users to choose the weight of each atomic feature. Given two aggregate feature elements E_1 and E_2:

$$\mathrm{Sim}(E_1, E_2) = f_{agg}\left(\mathrm{Sim}_i\left(e_i^1, e_i^2\right)\right) = \sum_{i=1..n} w_i \times \mathrm{Sim}_i\left(e_i^1, e_i^2\right) \in [0, 1]$$
$$\text{given } \sum_{i=1..n} w_i = 1 \quad \wedge \quad (w_{i=1..n}) \geq 0 \quad \wedge \quad \mathrm{Sim}_{i=1..n}(x, y) \in [0, 1] \tag{6}$$

where e_i^j is an atomic element describing feature *i* within aggregate element E_j, w_i is the weight of feature *i*, and Sim_i is the similarity according to feature *i*. For instance, the similarity between two patient EHRs described in Figure 8, considering aggregate feature elements made of atomic features *gender*, *pulse*, and *glycaemia*:

$$\mathrm{Sim}(E_1, E_2) = w_{gender} \times \mathrm{Sim}_{gender}(E_1, E_2) + w_{pulse} \times \mathrm{Sim}_{pulse}(E_1, E_2) + w_{glycaemia} \times \mathrm{Sim}_{glycaemia}(E_1, E_2)$$
$$= \frac{1}{3} \times \mathrm{Sim}_{gender}(\mathit{Female}, \mathit{Male}) + \frac{1}{3} \times \mathrm{Sim}_{pulse}(68, 75) + \frac{1}{3} \times \mathrm{Sim}_{glycaemia}(6.1, 6.6) = 0.653$$

We consider as reference $pulse_{max}$ = 170 bpm and $glycaemia_{max}$ =147 mmol/L for a middle aged human subject, to compute the atomic similarity functions accordingly[1] (cf. Table 1).

Table 1. Sample atomic element and feature vector similarity measures

Scalar values similarity	Comparing two scalar values x_i and x_j: $\mathrm{Sim}(x_i, x_j) = 1 - \frac{\lvert x_i - x_j \rvert}{x_{max}} \in [0, 1]$ where x_{max} is the maximum value from the reference dataset from which the values were sampled.	**(1)**
Date/Time stamps similarity	Comparing two date/time stamps x_i and x_j: $\mathrm{Sim}(x_i, x_j) = 1 - \frac{\lvert (x_i + x_{min}) - (x_j + x_{min}) \rvert}{\lvert x_{max} + x_{min} \rvert} \in [0, 1]$ where x_{max} and x_{min} are the maximum and minimum values from the reference dataset from which the date/time values were sampled.	**(2)**
Boolean values similarity	Comparing two Boolean values x_i and x_j: $\mathrm{Sim}(x_i, x_j) = x_i \wedge x_j$	**(3)**
String values similarity	Comparing two string values syntactically x_i and x_j: $\mathrm{Sim}(x_i, x_j) = 1 - \frac{EditDistance(x_i, x_j)}{\lvert x_i \rvert + \lvert x_j \rvert} \in [0, 1]$	**(4)**
Feature vectors similarity	Comparing two feature vectors V_i and V_j: $\mathrm{Sim}(V_i, V_j) = \frac{1}{n} \sum_{k=1}^{n} Sim(x_k^i, x_k^j)$	**(5)**

3.3. Data Clustering

In this study, we use the well-known Unweighted Pair-Group Method with Arithmetic mean (UPGMA) average link hierarchical clustering method [12, 15], although any form of hierarchical clustering can be utilized. Given *n* data points, we construct a fully connected graph *G* with *n* nodes and $\frac{n \times (n-1)}{2}$ weighted edges. The weight of an edge corresponds to the similarity (distance) between the connected nodes. We adopt an agglomerative clustering approach where each node in the connected graph initially

[1] *Gender* is modeled as a Boolean attribute, where *female* and *male* values are represented *true* (1) and *false* (0) respectively. We do not consider other gender types (e.g., transgender or gender neutral) since they do not exist within our patient data.

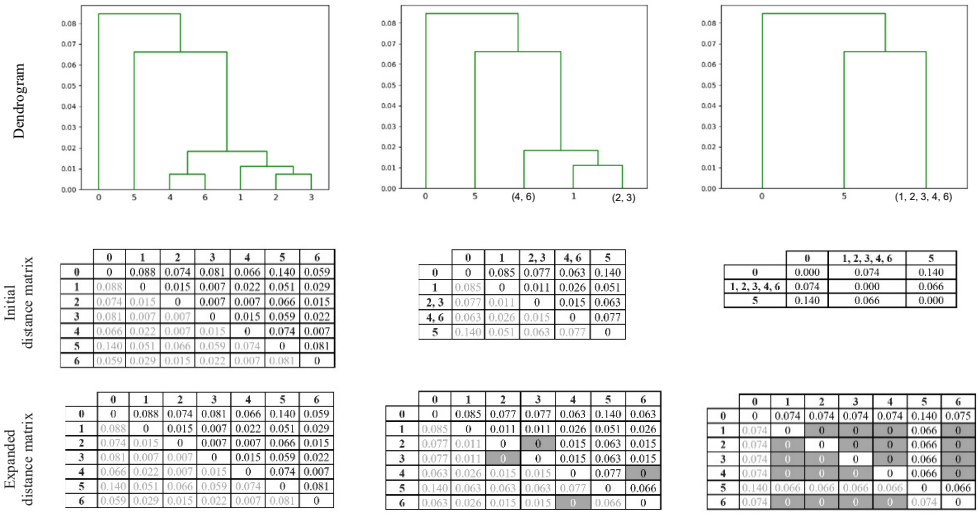
represents an individual cluster, where the similarity between the clusters is computed as the average of all similarities between their constituent edges. Figures 10 and 11 show the dendrograms and corresponding distance matrices produced for a sample dataset of 7 patient EHRs, clustered accordingly to the *Glycaemia* and *LDL* features[1] respectively (cf. experiments in Section 4).

3.4. Data Zooming

After performing the clustering process on the selected features and producing the resulting dendrogram structures, the tool recommends the best zooming level to display the dendrograms, according to a combined zooming score highlighting: i) the maximum similarity between the dendrograms, and ii) the minimal granularity for both dendrograms. More formally, given two dendrograms $dend_1$ and $dend_2$:

$$\text{zoomScore}(\text{dend}_1, \text{dend}_2) = \alpha \times \text{Sim}(\text{dend}_1, \text{dend}_2) + \beta \times (1 - \text{Gran}(\text{dend}_1, \text{dend}_2)) \quad \in [0, 1] \qquad \textbf{(7)}$$

where $\alpha, \beta \in [0, 1]$, $\alpha + \beta = 1$, $\text{Sim}(dend_1, dend_2) \in [0, 1]$, and $\text{Gran}(dend_1, dend_2) \in [0, 1]$. Similarly to the element aggregation measure mentioned in Section 3.2, we make use of the *weighted sum* function since it allows users to emphasize dendrogram similarity versus granularity according to their needs.



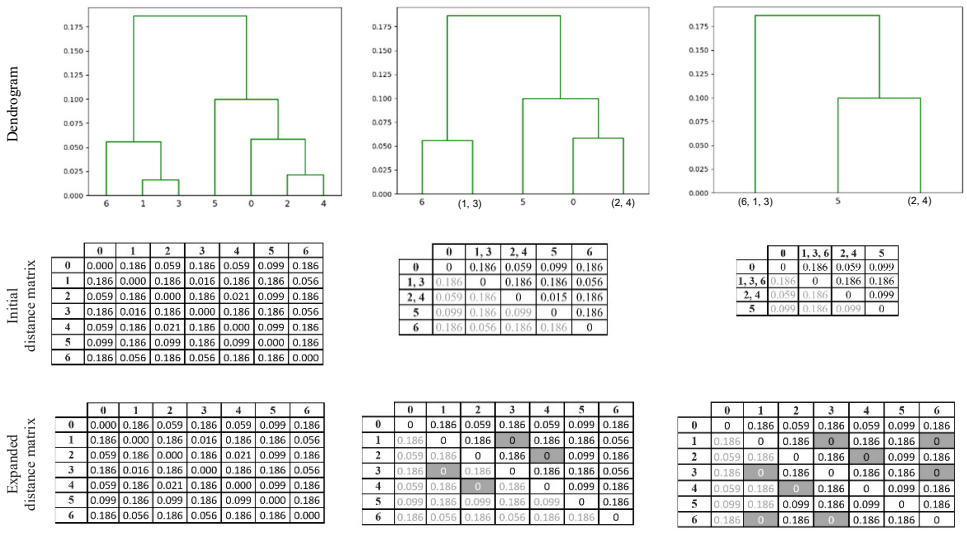
a. Zoom level = 7 (showing all data points)

b. Zoom level = 5 (with clusters (4, 6) and (2, 3))

c. Zoom level = 3 (with cluster (1,2,3,4,6))

Figure 9. Dendrograms produced for 7 patient EHRs clustered following the *Glycaemia* feature, with their distance matrices

[1] *Glycaemia* refers to the level of glucose in the patient's blood. LDL is commonly referred to as the "bad" cholesterol since it collects in the blood vessel walls.



a. Zoom level = 7 (showing all data points)

b. Zoom level = 5 (with clusters (1, 3) and (2, 4))

c. Zoom level = 3 (with clusters (1, 3, 6) and (2, 4))

Figure 10. Dendrograms produced for 7 patient EHRs clustered following the *LDL* feature, with their distance matrices

The zooming algorithm is shown in Figure 11. It accepts as input two sets of dendrograms produced for both features being compared including all zooming levels. It then computes the zooming score for each pair of dendrograms in both sets (lines 3-5) and identifies the pair which maximize the zooming score (lines 6-9).

Algorithm 1 – Duplicate Zooming
Input: DendSet1, DendSet2
Output: $dend_1$, $dend_2$

```
Begin
 1    maxZoomScore ← 0
 2    optimalZoomIndices ← {0, 0}
 3    For each dend_i ∈ DendSet1
 4        For each dend_j ∈ DendSet2
 5            if (maxZoomScore < zoomScore(dend_i, dend_j)) then
 6                maxZoomScore ← zoomScore(dend_i, dend_j)
 7                optimalZoomIndices ← {i, j}
 8    dend_1 ← dend_i
 9    dend_2 ← dend_j
 10   Return {dend_1, dend_2}
End
```

Figure 9. Pseudo code of our dendrogram zooming algorithm

3.4.1. Dendrogram Similarity

We evaluate the similarity between two dendrograms using their expanded distance matrices. The distance between a data point *x* and a cluster *Y* in the initial matrix, is represented as a replication of the same distance value between *x* and every data point

$y \in Y$ in the expanded matrix. We adopt the expanded distance matrices to maintain identical dimensionalities for both matrices being compared, regardless of hierarchical clustering (zooming) level (cf. Figures 9 and 10). We adopt normalized Manhattan distance to compute the similarity between a pair of data points, yet other vector similarity measures can be used. More formally:

$$\text{Sim}(\text{dend}_1, \text{dend}_2) = 1 - \text{Dist}(\text{dend}_1, \text{dend}_2) \;/\; \text{Dist}(\text{dend}_1, \text{dend}_2) = \frac{\sum_{i,j} |m_{i,j} - n_{i,j}|}{\sum_{i,j} |m_{i,j} + n_{i,j}|} \in [0, 1] \quad \textbf{(8)}$$

where $m_{i,j}$ is the distance entry in the distance matrix corresponding to $dend_1$, and $n_{i,j}$ is the distance entry in the distance matrix corresponding to $dend_2$. Table 2 shows the pair-wise similarity scores between pairs of dendrograms produced following our *Glycemia* vs *LDL* running example. An entry at position (4, 5) in the similarity matrix represents the similarity score between the dendrogram of zooming level =4 for *Glycemia* and the dendrogram of zooming level =5 for *LDL*.

Table 2. Similarity matrix for Glycemia vs LDL dendrograms

Dend#	1	2	3	4	5	6	7
1	NaN	0	0	0	0	0	0
2	0	0.1851	0.2226	0.2885	0.2787	0.2768	0.2755
3	0	0.2944	0.3971	0.4509	0.4369	0.4343	0.4324
4	0	0.3189	0.4162	0.4678	0.4735	0.4806	0.4786
5	0	0.3235	0.4199	0.4709	0.4766	0.4837	0.4874
6	0	0.3273	0.4232	0.4741	0.4796	0.4866	0.4904
7	0	0.3312	0.4265	0.4771	0.4825	0.4896	**0.4933**

Table 3. Granularity matrix for Glycemia vs LDL dendrograms

1	2	3	4	5	6	7
0	0.0833	0.1667	0.2500	0.3333	0.4167	0.5000
0.0833	0.1667	0.2500	0.3333	0.4167	0.5000	0.5833
0.1667	0.2500	0.3333	0.4167	0.5000	0.5833	0.6667
0.2500	0.3333	0.4167	0.5000	0.5833	0.6667	0.7500
0.3333	0.4167	0.5000	0.5833	0.6667	0.7500	0.8333
0.4167	0.5000	0.5833	0.6667	0.7500	0.8333	0.9167
0.5	0.5833	0.6667	0.7500	0.8333	0.9167	1

3.4.2. Dendrogram Granularity

In addition to maximum dendrogram similarity, our solution recommends the best zooming level to display the dendrograms with the minimum granularity, i.e., minimum amount of information details presented to the user. More formally:

$$\text{Gran}(\text{dend}_1, \text{dend}_2) = \alpha \times \text{Gran}(\text{dend}_1) + \beta \times \text{Gran}(\text{dend}_2) \;/\; \text{Gran}(\text{dend}_i) = \frac{\#\text{leaf nodes}(\text{dend}_i) - 1}{\#\text{data points}(\text{dend}_i) - 1} \in [0, 1] \quad \textbf{(9)}$$

where $\alpha, \beta \in [0, 1]$, and $\alpha + \beta = 1$. A granularity score =1 means that the dendrogram is fully zoomed-in, showing the maximum number of nodes (i.e., the maximum amount of information details). A granularity score = 0 means that the dendrogram is fully zoomed-out, showing the minimum number of nodes =1 (i.e., the root node only, highlighting minimum details).

Table 3 shows the pair-wise granularity scores between all pairs of dendrograms from our *Glycemia* vs *LDL* running example, considering equal weights for individual granularity scores ($\alpha = \beta = 0.5$). The granularity score between the dendrograms at the lowest levels is =0. The granularity score between the dendrograms at the highest levels =1. The granularity score increases with the zoom level, and decreases accordingly. Following several experimental runs (cf. Section 4), we assign a weight $\alpha = 0.8$ for the dendrogram similarity score and $\beta = 0.2$ for the dendrogram granularity score (other weight configurations can be considered). Results show that the *Glycemia* dendrogram

of level =3 and the *LDL* dendrogrma of level =4 produce the maximum zoomScore value =0.4774, and thus will be returned by the system as the best zooming level to display the dendrograms (cf. Figure 12).

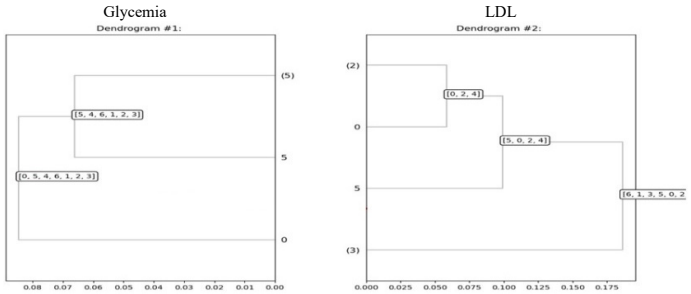


Figure 12. The best dendrogram zooming levels for *Glycemia* at level 3 versus *LDL* at level 4

3.5. Node Connections

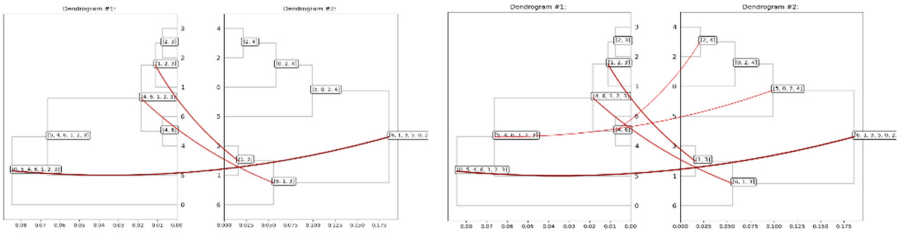
Following the identification of the best zooming level among the paired dendrograms, the remaining step is to connect the internal nodes of the dendrograms in order to highlight their correlation. To achieve this, we compute dendrogram internal node similarity as the similarity between the corresponding clusters, represented as bags of data points. We utilize Jaccard similarity, yet other set similarity measures can be used (e.g., Intersection, Dice). More formally, considering two dendrograms $dend_1$ and $dend_2$, and two internal nodes $x_i \in dend_1$ and $y_j \in dend_j$ being compared:

$$\mathrm{Sim}(x_i, y_j) = \frac{|\,\mathrm{cluster}(x_i) \cap \mathrm{cluster}(y_j)\,|}{|\,\mathrm{cluster}(x_i) \cup \mathrm{cluster}(y_j)\,|} \in [0, 1] \quad \textbf{(10)}$$

where $cluster(x_i)$ and $cluster(y_j)$ are the clusters represented by nodes x_i and y_j. Consequently, we use the transportation optimization problem, e.g., [21, 22], to match the related internal nodes from both dendrograms. The transportation problem seeks to associate a number of supply centers m (sources) with a number of demand centers n (destinations) to optimize supply delivery. We consider the internal nodes of the first dendrogram to be the supply centers, and the internal nodes of the second dendrogram to be the demand centers. Considering two dendrograms with m and n internal nodes respectively, we build an $m \times n$ matrix where the rows represent the internal nodes of the first dendrogram and the columns represent the internal nodes of the second dendrogram. Each entry (i, j) provides the similarity between internal node x_i from the first dendrogram, and internal node y_j from the second dendrogram. Consider for instance the fully zoomed-in visualization of *Glycemia* vs *LDL* in Figure 13, with zoom level =7 for both dendrograms. Hence, we have m-1 = n-1 = 6, resulting in a 6×6 pairwise internal node similarity matrix shown in Table 4.

Once the internal node similarity matrix is produced, we start by matching the nodes together using the transportation problem's *minimum* (*least*) *cost method* widely adopted in the literature, e.g., [21, 22] (other approaches can be used such as *penalty-based* or *correction-based* methods [3]). We compute cost as the inverse of similarity, and hence we seek to minimize the cost among the matching nodes (cf. Table 4.b).

Once all the internal node connections have been established, the system displays all the connections having a similarity score greater than or equal to a (user or system-defined) threshold. Figure 13.a shows the internal node connections having similarity scores above 0.5 (sharing more than 50% similarity). Figure 13.b shows more internal connections after lowering the similarity threshold to 0.3. Also, the thickness of the node connections is defined proportionally to their similarity, where thicker connections highlight more similar nodes.



a. Node connections with a threshold value of 0.5 **b.** Node connections with a threshold value of 0.3

Figure 13. Full zoomed-in visualization of Glycemia vs LDL

Table 4. Internal nodes similarity matrix for full zoomed-in visualization of Glycemia vs LDL

a. Initial similarity matrix

Glycemia dendrogram internal clusters	*LDL* dendrogram internal clusters: (1, 3)	(2, 4)	(6, 1, 3)	(0, 2, 4)	(5, 0, 2, 4)	(6, 1, 3, 5, 0, 2, 4)
(4, 6)	0	0.3333	0.25	0.25	0.2	0.2857
(2, 3)	0.3333	0.3333	0.25	0.25	0.2	0.2857
(1, 2, 3)	0.6667	0.25	0.5	0.2	0.1667	0.4288
(4, 6, 1, 2, 3)	0.4	0.4	0.6	0.3333	0.2857	0.7143
(5, 4, 6, 1, 2, 3)	0.3333	0.3333	0.5	0.2857	0.4286	0.8571
(0, 5, 4, 6, 1, 2, 3)	0.2857	0.2857	0.4286	0.4286	0.5714	1

b. Result of the transportation problem's minimum cost method, where the order of the iteratively selected cells is shown in subscript

Glycemia dendrogram internal clusters	*LDL* dendrogram internal clusters: (1, 3)	(2, 4)	(6, 1, 3)	(0, 2, 4)	(5, 0, 2, 4)	(6, 1, 3, 5, 0, 2, 4)
(4, 6)	0	0.3333	0.25	**0.25$_6$**	0.2	0.2857
(2, 3)	0.3333	**0.3333$_5$**	0.25	0.25	0.2	0.2857
(1, 2, 3)	**0.6667$_2$**	0.25	0.5	0.2	0.1667	0.4288
(4, 6, 1, 2, 3)	0.4	0.4	**0.6$_3$**	0.3333	0.2857	0.7143
(5, 4, 6, 1, 2, 3)	0.3333	0.3333	0.5	0.2857	**0.4286$_4$**	0.8571
(0, 5, 4, 6, 1, 2, 3)	0.2857	0.2857	0.4286	0.4286	0.5714	**1$_1$**

4. Experimental Evaluation

We have implemented our tool using the Python programming language. We perform text preprocessing and feature extraction using *NLTK*, matrix computations using *NumPy*, clustering and dendrogram building using *SciPy*, dendrogram visualization using *MatplotLib*, and GUI functionalities using *Tkinter*. The tool is available online[1].

4.1. EHR Case Study

We used a sample dataset of 114 EHRs of patients who suffer from migraine disorder, obtained from a private medical clinic where all EHRs were vetted by Dr. Sola Aoun Bahous, M.D. and professor in the department of internal medicine at LAU Rizk hospital. Sample EHR extracts are shown in Figure 7. We conducted tests to visualize correlated and uncorrelated features and compare the results with existing tools.

[1] http://sigappfr.acm.org/Projects/MirroredDendrograms/

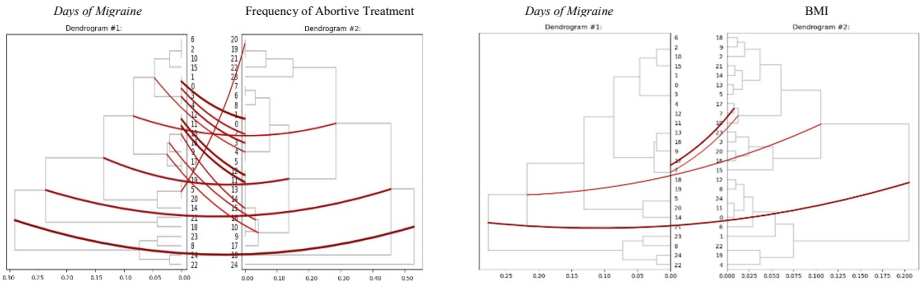
4.1.1. Feature Correlation

We compare: i) a pair of correlated features: *days of migraine* and *frequency of abortive treatment* having average correlation pcc[1] = 0.5882, and ii) and a pair of less correlated features: *days of migraine* and *BMI*[2] having average pcc = 0.1556. Table 5 shows a subset of the data, and the corresponding mirrored dendrogram visualizations are shown in Figure 14. A larger subset is visualized in Figure 15 with varying zooming levels.

Table 5. Sample EHR features for a subset of 25 patients

Days of Migraine	10	10	12	10	10	20	12	16	28	14	12	15	15	15	18	13	15	16	16	20	20	4	25	30	26
Frequency of Abortive Treatment	10	10	10	10	10	10	12	12	13	14	14	15	15	15	15	15	15	16	16	20	20	20	22	25	30
BMI	27.09	25.71	21.09	20.02	28.62	22.98	26.61	23.71	27.68	21.51	24.21	27.34	27.99	22.76	22.22	18.25	19.59	23.95	21.87	29.74	19.29	22.49	29.75	20.34	27.34

Based on Figure 14, we highlight the following: i) the mirrored dendrograms in Figure 14.a show similar structures with many connected nodes, reflecting high feature correlation, ii) the mirrored dendrograms in Figure 14.b show less similar structures with only four pairs of connected nodes, reflecting low feature correlation. Similar observations are obtained with the larger data subset in Figure 15, where the high and low correlations are reflected in Figure 15.a and b respectively. We obtain similar observations using different zooming levels in Figure 15.c and d.



a. Correlated features **b.** Less correlated features

Figure 14. Mirrored dendrogram visualizations for two pairs of sample EHR features considering a subset of 25 patients from Table 5, shown according to the best zooming levels identified by the tool, with node connection threshold = 0.5

4.1.2. Comparison with Alternative Solutions

In addition, we compare our tool with two alternative visualizations: tanglegram and cluster heatmap. We use the sample dataset and pairs of EHR features from the previous example. Results are shown in Figure 16. While designed to describe the correlations between pairs of dendrograms, yet both tanglegram and cluster heatmap compare dendrograms according to their leaf node mapping, and do not visualize the similarities within the structures themselves. This can be misleading since two dendrograms can have different internal structures, while their leaf nodes are presented in a matching

[1] Pearson Correlation Coefficient
[2] Body Mass Index

order, and vice versa. This is the case in Figure 16 where both the highly correlated features in Figures 16.a and c and the less correlated features in Figures 16.b and d produce similar tanglegram and cluster heatmap visualizations respectively, making it difficult to judge the correlations between the compared features. Different from tanglegram and cluster heatmap, our tool i) computes the similarity between dendrogram structures and maps their internal nodes to describe their structure relationships, ii) allows to zoom-in and out of the data to show their relationships at different granularity levels (compared with existing static solutions), and iii) identifies the best zooming level between the two dendrograms, highlighting the maximum correlation with the minimal amount of details presented to the user.

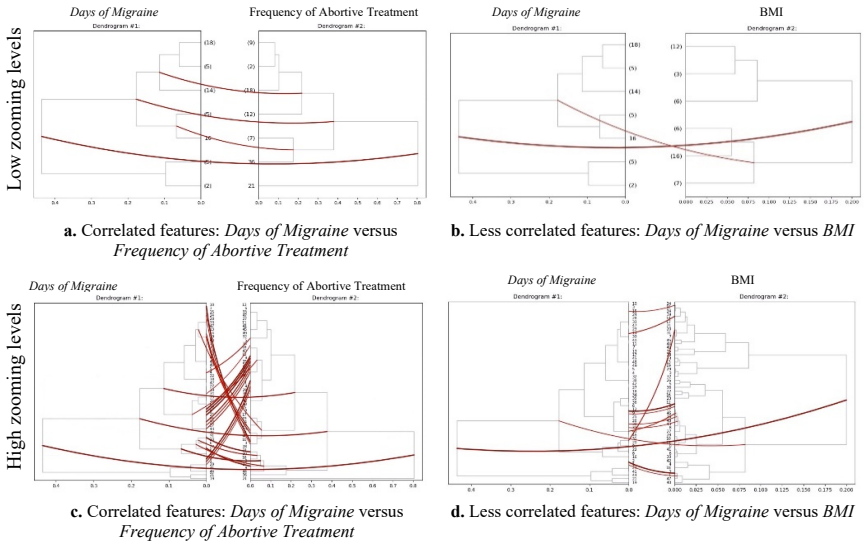
4.2. User Study

Since our work involves visualizations perceived by users, we acquired and evaluated the feedback from human testers to assess the quality of our visualization tool. For this purpose, we created an online survey[1] considering five evaluation criteria: i) feature correlation visualization, ii) default zooming levels, iii) zooming in and out actions, iv) tool's interactive functionalities, v) comparison with existing solutions (cf. Table 6). A total of 20 were invited to contribute to the experiment, where they independently rated every evaluation criterion on an integer scale from 1 to 10 (i.e., from *highly dissatisfied* to *highly satisfied*). Testers were undergraduate and graduate engineering students, as well as junior and senior engineers with background in data science, business analytics, computer science, or computer engineering (cf. Figure 17). An invitation email was shared by the authors and broadcast to their undergraduate and graduate engineering students and alumni. The first 20 testers who accepted the invitation volunteered to conduct the survey and did not receive any compensation. Testers were initially shown a demo of the mirrored dendrogram, tanglegram, and cluster heatmap tools, providing them with sample visualizations for every tool. Testers were also invited to use the tools on three small data samples provided by the authors, to familiarize with their visualizations and functionality, including the inner node connections and zooming functionalities provided by mirrored dendrograms.

Table 6. Visualization tool's evaluation criteria

Criterion	Description	Evaluation question
1. Feature correlation visualization	Ability of the tool to allow users to visually distinguish between highly correlated features and loosely correlated features, when mirrored against each other.	How satisfied are you with the feature correlation visualization of the tool?
2. Default zooming levels	Quality of the default zooming levels suggested by the tool, highlighting the maximum correlation with the minimal amount of details presented to the user.	How satisfied are you with the tool's default zooming levels?
3. Zooming in and out actions	How efficient it is to zoom in and out of the data, and navigate up and down the dendrogram hierarchies.	How satisfied are you with the zooming actions of the tool?
4. Tool's interactivity	Capacity of tool to provide interactive functionalities, including parameter settings, similarity thresholds, node and edge visualizations and coloring, among others.	How satisfied are you with the tool's interactive functionalities?
5. Comparing visualization quality with tools	Quality of the tool's visualization compared with existing solutions: namely tanglegram and cluster heatmap.	How satisfied are you with the tool's visualization quality compared with existing solutions?

[1] Available at: https://github.com/akf98/mirrored-dendrogram-tool



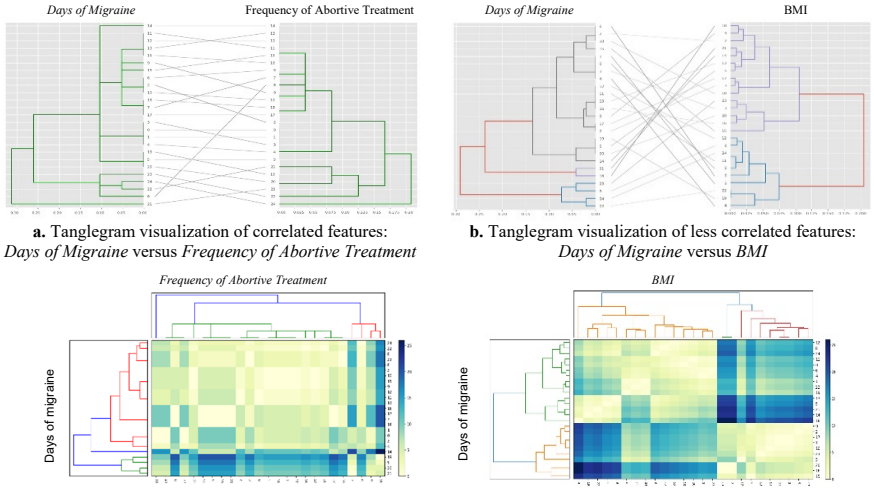
a. Correlated features: *Days of Migraine* versus *Frequency of Abortive Treatment*

b. Less correlated features: *Days of Migraine* versus *BMI*

c. Correlated features: *Days of Migraine* versus *Frequency of Abortive Treatment*

d. Less correlated features: *Days of Migraine* versus *BMI*

Figure 15. Mirrored dendrogram visualizations for two pairs of sample EHR features considering a subset of 50 patients, shown according to the best zooming levels identified by the tool, with node connection threshold = 0.5



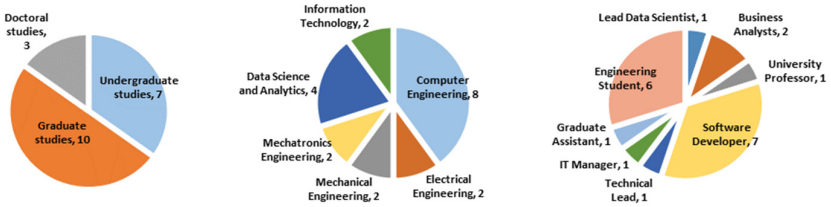
a. Tanglegram visualization of correlated features: *Days of Migraine* versus *Frequency of Abortive Treatment*

b. Tanglegram visualization of less correlated features: *Days of Migraine* versus *BMI*

c. Cluster heatmap visualization of correlated features: *Days of Migraine* versus *Frequency of Abortive Treatment*

d. Cluster visualization of less correlated features: *Days of Migraine* versus *BMI*

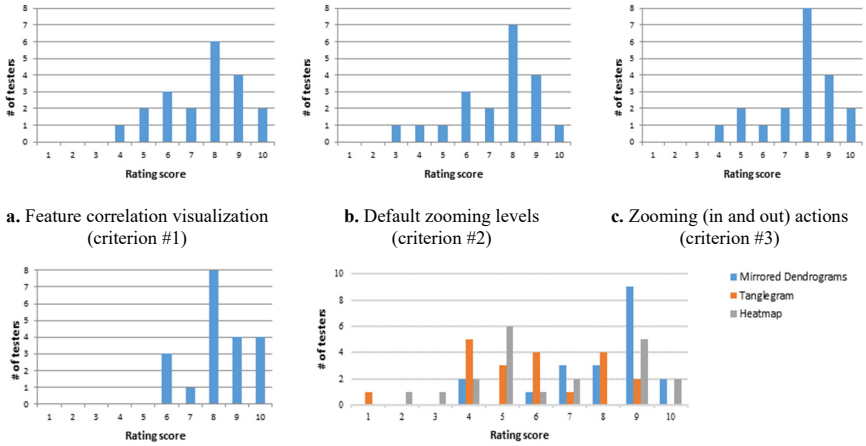
Figure 16. Tanglegram and cluster heatmap for pairs of sample EHR features from Table 5



a. Educational level **b.** Field of study **c.** Profession

Figure 17. Testers' education level, and field of study, and professions

Results in Figure 18 and 19 show the compiled tester ratings, and the average rating scores aggregated for every criterion. We summarize the results: i) *Feature correlation visualization:* Results show that 68.5% of the testers gave this criterion scores ≥ 7/10, achieving an average score of 7.5/10 (stdev = 1.7) ; ii) *Default zooming*: Results show that 68.4% of the testers gave this criterion scores ≥ 7/10, achieving an average score of 7.3/10 (stdev = 1.8) ; iii) *Zooming actions*: Results show that 78.9% of the testers gave this criterion scores ≥ 7/10, achieving an average score of 7.7/10 (stdev = 1.6) ; iv) *Tools' interactivity*: Results show that 65.8% of the testers gave this criterion scores ≥ 7/10, achieving an average score of 8.3/10 (stdev = 1.3) ; v) *Comparative evaluation*: Results show that 84.2% of the testers gave the mirrored dendrograms rating scores ≥ 7/10, compared with 36.8% and 47.4% for tanglegram and cluster heatmaps respectively. The mirrored dendrograms achieved an average rating of 8 (stdev = 1.7), compared with 5.85 (stdev = 2) and 6.4 (stdev = 2.5) for tanglegram and cluster heatmaps respectively.



a. Feature correlation visualization (criterion #1)

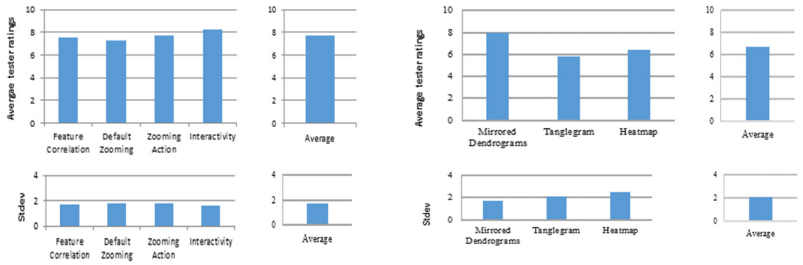
b. Default zooming levels (criterion #2)

c. Zooming (in and out) actions (criterion #3)

d. Tool's interactivity (criterion #4)

e. Comparing visualization quality with existing tools (criterion #5)

Figure 18. Tester rating scores for every evaluation criterion



a. Average ratings for individual criteria **b.** Average ratings for comparative evaluation

Figure 19. Average tester ratings for all evaluation criteria

Results show that most testers are satisfied with the visualization tool: i) describing feature correlations, ii) suggesting a default zooming level to compromise between maximum correlation and minimal amount of details presented to the user, iii) zooming-in and out the data to visualize cluster hierarchies at different levels of details, and vi) providing improved visualizations compared with existing solutions.

5. Conclusion

We introduce an unsupervised feature-based tool for interactive data visualization titled "mirrored dendrograms". It accepts as input semi-structured EHRs and allows the user to select the target features to be mapped against each other. It produces a dendrogram structure for each combination of features, connecting the data's internal nodes to describe their relationships. The user can zoom-in and out of the data to show their relationships at different granularity. The tool also identifies the best zooming level which highlights the maximum correlation with the minimal amount of details presented to the user. Empirical results highlight the tool's performance. We are currently extending the tool to consider the time dimension, producing a 3D visualization where data belonging to the same timestamp will be clustered and presented on a plane related to the timestamp. This is crucial to correlate time-stamped social media data (e.g., describing social event correlations [1, 25]) and sensor network data (e.g., describing network event correlations [10, 11]).

References

[1] Abebe M., et al., *Generic Metadata Representation Framework for Social-based Event Detection, Description, and Linkage.* Knowledge Based Systems 2020. 188.

[2] Ahmad A. and Khan S., *Survey of State-of-the-Art Mixed Data Clustering Algorithms.* IEEE Access 2019. 7: 31883-31902.

[3] Biswas A., et al., *A Study of Multi-Objective Restricted Multi-Item Fixed Charge Transportation Problem considering Different Types of Demands.* Applied Soft Computing, 2022. 118:108501.

[4] Bok J., et al., *Augmenting Parallel Coordinates Plots With Color-Coded Stacked Histograms.* IEEE Trans. Vis. Comput. Graph, 2022. 28(7): 2563-2576.

[5] Britzolakis A., et al., *AthPPA: A Data Visualization Tool for Identifying Political Popularity over Twitter.* Information journal, 2021. 12(8): 312.

[6] Buchin K., et al., *Drawing (Complete) Binary Tanglegrams - Hardness, Approximation, Fixed-Parameter Tractability.* Algorithmica, 2012. 62(1-2): 309-332.

[7] Chen C., et al., *Detecting the Association of Health Problems in Consumer-level Medical Text.* Journal of Information science, 2018. 44(1): 3-14.
[8] De Vienne D., *Tanglegrams are Misleading for Visual Evaluation of Tree Congruence.* Molecular Biology and Evolution, 2019. 36(1): 174-176, doi:10.1093/molbev/msy196.
[9] Dwyer T., *Scalable, Versatile and Simple Constrained Graph Layout.* Comput Graph Forum, 2009. 28(3):991–8.
[10] Ebrahimi D., et al., *Data Collection in Wireless Sensor Networks Using UAV and Compressive Data Gathering.* GLOBECOM, 2018. pp. 1-7.
[11] Ebrahimi D., et al., *UAV-Aided Projection-based Compressive Data Gathering in Wireless Sensor Networks.* IEEE Internet Things journal, 2019. 6(2): 1893-1905.
[12] Edwards R., *UPGMA Worked Example.* Edwards Lab, University of New South Whales, Australia, 2016. http://www.slimsuite.unsw.edu.au/teaching/upgma/.
[13] Engle S., et al., *Unboxing Cluster Heatmaps.* Proceedings of the Symposium on Biological Data Visualization (VIS'17), 2017. 18(S-2):63:1-63:15.
[14] Galili T., et al., *Heatmaply: an R Package for Creating Interactive Cluster Heatmaps for Online Publishing.* Bioinformatics, 2018. 34(9):1600-1602.
[15] Halkidi M.;Batistakis Y. and Vazirgiannis M., *Clustering Algorithms and Validity Measures.* Inter. Conf. on Scientific and Statistical DB Management (SSDBM), 2001, 3-22.
[16] Johansson J., et al., *3-Dimensional Display for Clustered Multi-Relational Parallel Coordinates.* Inter. Conf. on Information Visualisation, 2005. pp. 188-193.
[17] NCSS Statistical Software, *Clustered Heatmaps.* 2022. Ch. 450, pp. 1-12, http://ncss.com.
[18] Nohno K., et al., *Spectral-Based Contractible Parallel Coordinates.* Inter. Conf. on Information Visualization, Paris, France. , 2014. pp. 7-12.
[19] Raj. J., *7 Ways Data Visualization Can Improve Sales and Marketing Alignment.* In Intellectyx, 2019. https://www.intellectyx.com/blog/ways-data-visualization-can-improve-sales-and-marketing-alignment/.
[20] Sakai R., et al., *Modular Leaf Ordering Methods for Dendrogram Representations in R.* F1000Research, 2014. 3(177).
[21] Salazar R., *Operations Research with R - Transportation Problem.* Towards Data Science, 2019. https://towardsdatascience.com/operations-research-in-r-transportation-problem-1df59961b2ad.
[22] Salloum G. and Tekli T., *Automated and Personalized Meal Plan Generation and Relevance Scoring using a Multi-Factor Adaptation of the Transportation Problem.* Soft Computing, 2022. 26(5): 2561-2585.
[23] Simpao A., et al., *A Review of Analytics and Clinical Informatics in Health Care.* J. of Med. Sys., 2014. 38(4), 1-7.
[24] Stasko J. and Zhang E., *Focus+ Context Display and Navigation Techniques for Enhancing Radial, Space-filling Hierarchy Visualizations.* IEEE Symposium on Information Visualization, 2000. p. 57–65.
[25] Taddesse F.G., et al., *Semantic-based Merging of RSS Items.* World Wide Web journal, 2010. 13(1-2): 169-207.
[26] Tekli J., et al., *Minimizing User Effort in XML Grammar Matching.* Information Sciences J., 2012. 210:1-40.
[27] Tekli J., et al., *(k, l)-Clustering for Transactional Data Streams Anonymization.* Information Security Practice and Experience, 2018. pp. 544-556.
[28] Tekli J., *An Overview of Cluster-based Image Search Result Organization: Background, Techniques, and Ongoing Challenges.* Knowl. Inf. Syst., 2022. 64(3): 589-642.
[29] Wang W., et al., *Visualization of Large Hierarchical Data by Circle Packing.* Conference on Human Factors in Computing Systems, 2006. p. 517-20.
[30] Weinstein J., *A Postgenomic Visual Icon.* Science jounal 2008. 319(5871):1772–3.
[31] Zhonghua Y. and Lingda W., *3D-Parallel Coordinates: Visualization for Time Varying Multidimensional Data.* Inter. ICSESS'16 Conf., 2016. doi:10.1109/ICSESS.2016.7883153.