Enhanced Literature Review Visualization: a Novel Sorted Stream Graphs with Integrated Word Elements

Vinh Nguyen The and Phung Trung-Nghia

September 25, 2024

Enhanced Literature Review Visualization: A Novel Sorted Stream Graphs with Integrated Word Elements

The-Vinh Nguyen[0000−0002−1300−3943], Trung-Nghia Phung[0000−0003−0075−3427]

Thai Nguyen University of Information and Communication Technology, Thai Nguyen, Vietnam
vinhnt@ictu.edu.vn

Abstract. Extracting meaningful insights from temporal data through visualization plays a crucial in decision-making process. Conventional visualization methods such as stream graphs and stacked area charts suffer from clarity when applying to show trends of cross-categories over time. To alleviate this limitation, we propose the Sorted Stream Graph with Embedded Word Elements (SSGEW), an innovative approach that sorts stream segments and incorporates embedded word elements into a stream graph. Through an authored algorithmic development, our method enhances the arrangement of data categories and strategically places words to improve data exploration. We compare our visual design with two traditional techniques and demonstrate our approach through a case study on the evolution of automatically generated data visualization from 2017 to 2024. Our results show a clear distinction between SSGEW and the two other designs, especially when there are many fluctuations in cross-categories. Future work could be focusing on refining the design to overcome occlusion.

Keywords: Data Visualization · Sorted Stream Graph · Temporal Data Analysis · Embedded Words · Generative AI.

1 Introduction

In today's data-driven era, extracting meaningful insights from temporal data through visualization plays a crucial in decision-making process. Edward Tufte [12], Jeffrey Heer [3]; Tamara Munzner [9], Ben Shneiderman [11] and Hadley Wickham [13] are among the top contributors to this active domain. Many graphical representations for data such as bar charts, histograms, line graphs, area plots, and stream graphs [13, 11, 10] are available to support users building visual format. Among these, stream graph gains its popularity due to its inherent characteristics. Derived from the Stacked Area Charts, the stream graph makes a further refinement by repositioning the central baseline and symmetrically organizing data around it [4]. This intention is to create a good-looking visual design with smooth data flow over time.

Many efforts have been made to demonstrate the application of stream graphs in the literature (e.g., [1, 7, 2, 6]). For example, to illustrate World War I's trends [1], to analyze the progression of topics on social media platforms [7] or to integrate Word Cloud elements [6]. Along with advantages, the stream graph also contains some challenges. One such limitation is the difficulty in comparing cross-category information at specific time points or intervals. This constraint can pose difficulties for users aiming to analyze and contrast the development trajectories of multiple subjects over a period. To the best of our knowledge, no prior research has specifically examined this barrier.

To fill this gap and alleviates previous issue, the current study proposes an innovative idea that enables multi-categorical comparison of trends in data over time called sorted stream graph. Furthermore, embedded word elements are also integrated into the stream segment to get insights in each time period. Fig. 1 illustrates our work (C) compared to previous efforts (A) and (B). In this Figure, (A) is a regular stacked area chart where each area is stacked on top of each other. (B) is a stream graph with the base positioned at the center. This integrated approach enables data analysts to capture patterns of the dynamic data variation as well as their corresponding linguistic implications. Thus, our study provided the following contributions:

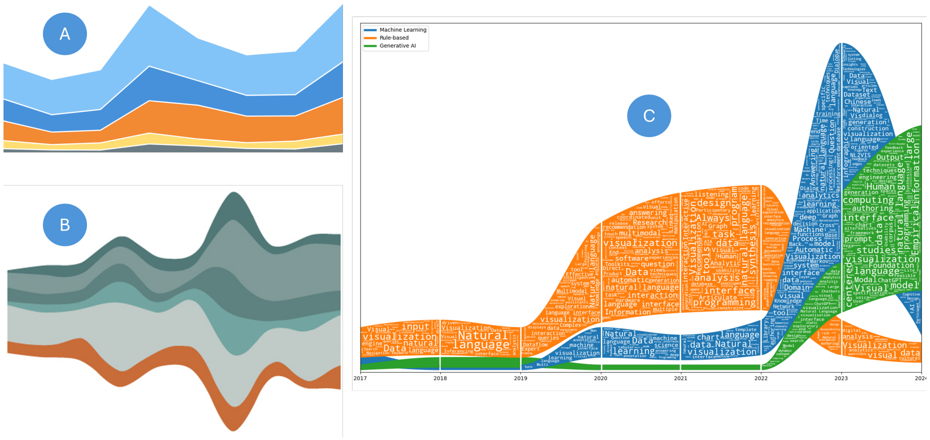

Fig. 1. (A) Stacked Area Chart, (B) Stream Graph, (C) Our work: Sorted Stream Graph

- A unique idea to segregate segments within a stream graph based on categories/topics of interest
- The development of authored algorithms that transform this conceptual idea to implementation.
- The practical demonstration of the sorted stream graph as a new visual design.

2 Method

2.1 Data Collection and Preparation

We utilized a dataset ($\mathbf{D}$) comprising documents with attributes such as Year (y) - or any discrete ordinal value, Category (c), Title (t) of the document, and Keywords (k). Missing values in the Keywords's field were replaced with empty string ($k_i ='' $ if k_i is null). The unique years ($\mathbf{D}_y$) and Category ($\mathbf{D}_d$) were extracted from the dataset.

2.2 Aggregation and Sorting Algorithm

Aggregation:

We created a complete index ($\mathbf{I}$) for all possible Year-Category combination ($\mathbf{D}_y \times \mathbf{D}_c$) and grouped the dataset accordingly. For each combination, we aggregated the data to obtain the total number of records ($P(y,c)$) and concatenated keywords ($K(y,c)$)

$$\mathbf{I} = \{(y, c) \mid y \in \mathbf{D}_y, c \in \mathbf{D}_c\}$$

$$g(\mathbf{D}, y, c) = \left(\sum_{t \in \mathbf{D}|y,c} 1, \bigcup_{k \in \mathbf{D}|y,c} k \right)$$

This aggregation ensured that all time points have a value (even if it is not presented - then filled it with zero).

Algorithm 1: Sorting algorithm by time point

```
$\forall i \in \{0, 1, \ldots, n-1\}, \forall j \in \{0, 1, \ldots, m-1\}$ :
$a_i = [a^i_{k_1}, \ldots, a^i_{k_m}]$
$b_i = \{(k, a^i_k) \mid k = 0, \ldots, m-1\}$
$c_i = \text{sorted}(b_i, \lambda x : x[1])$
$d_i = [a^i_{k_{j_1}}, \ldots, a^i_{k_{j_m}}]$
$e_i = [j_1, \ldots, j_m]$
$f = []$
$\forall j \in \{0, \ldots, m-1\}$ :
if $j = 0$ and $a^i_{e_{ij}} \neq 0$ then
        $f.append(a^i_{e_{ij}})$
        $a^i_{e_{ij}} = \frac{a^i_{e_{ij}}}{2}$
else
        $p = a^i_{e_{i(j-1)}}$
        $r = \frac{f[j-1]}{2}$
        $q = a^i_{e_{ij}}$
        $f.append(q)$
        $a^i_{e_{ij}} = p + r + \frac{q}{2} + g$
end
$\forall k \in \{0, \ldots, m-1\}, h_k.append(a^i_k)$
```

Where:

n is the number of years, m is the number of category, a_i is the list of values for iteration i, b_i is the indexed list of values, c_i is the sorted indexed list, d_i is the sorted list of values, e_i is the list of original indices, f is the list of copied original values, p is the previous value, r is the radius (half of the previous copied original value), q is the current value, g is the gap, h_k is the line chart for index k, $f.append(x)$ denotes appending x to the list f

2.3 Smoothing and Interpolation

To smooth the trend lines over years, we used Piecewise Cubic Hermite Interpolating Polynomial (PCHIP) for interpolation [8]. For each method, we constructed category lines ($\mathbf{L}_c$) representing the total documents per year. These lines were smoothed to generate interpolated values ($P_{smooth}(x)$) and their bounds ($P_{lower}(x)$ and $P_{upper}(x)$) using the following formulas:

$$P_{smooth}(x) = \text{PCHIPInterpolator}(\mathbf{X}, \mathbf{L}_m)$$

$$P_{lower}(x) = P_{smooth}(x) - \alpha |P_{smooth}(x)|$$

$$P_{upper}(x) = P_{smooth}(x) + \alpha |P_{smooth}(x)|$$

where $\mathbf{X}$ represents the unique time points, and α is a smoothing parameter.

2.4 Overlaying graph segment by time point

Given:

- $\{y_i\} = \{y_1, y_2, \ldots, y_n\}$, where y_i are the sorted unique time points.
- $P_c = \{P_c(y_1), P_c(y_2), \ldots, P_c(y_n)\}$ represents the document counts for each category c.

For each interval $[y_i, y_{i+1}]$ where $i = 0, 1, \ldots, n-2$:

- Compute S_c, the sorted document counts for category c:

$$S_c = \text{sort}(P_c[y_{i+1}])$$

- Determine I_c, the indices of P_c after sorting:

$$I_c[j] = \text{index of } S_c[j] \text{ in } P_c$$

- Calculate the zorder Z_c for each $P_c(y_{i+1})$:

$$Z_c[j] = \text{index of } P_c(y_{i+1}) \text{ in } S_c$$

2.5 Plotting and Embedded Word Elements

For each year y_i in the dataset, we plotted the interpolated values and filled the area between the bounds to represent the variability in the data. These filled areas were used as masks for embedding text or words. The plotting process was conducted as follows:

1. The areas between the bounds were filled with colors from a predefined set $\mathbf{C}$. For each interval $[y_i, y_{i+1}]$:

$$\text{Fill}(y_i, y_{i+1}, \mathbf{C}_c, Z_c) = \{(x, y) \mid y \in [P_{c,\text{lower}}(x), P_{c,\text{upper}}(x)], x \in [y_i, y_{i+1}], Z_c, \mathbf{C}_c\}$$

2. These filled areas were extracted as masks for words. It is noted that we initialize the mask size equal to all filled areas (i.e., 1200x800). This is to ensure that the location of the mask remains its original position. For each year-category combination (y_i, c), the mask was defined as:

$$M(x, y) = \begin{cases} 1 & \text{if pixel } (x, y) \text{ is white} \\ 0 & \text{if pixel } (x, y) \text{ is not white} \end{cases}$$

$$(x_0, y_0) = \left(\frac{W}{2}, \frac{H}{2}\right)$$

and the parametric spiral equation for word placement is calculated as:

$$(x(t), y(t)) = (x_0 + at\cos(\omega t + \phi), y_0 + at\sin(\omega t + \phi))$$

where a is the scaling factor, ω is the angular frequency, and ϕ is the phase shift.

The collision detection and bounding box placement inside the mask is calculated as:

$$\text{collision}(B_i, M) = \sum_{(x,y)\in B_i} \left(M(x, y) \cdot \prod_{j<i} [1 - \chi_{B_j}(x, y)] \right) = 0$$

where χ_{B_j}(x, y) is the indicator function for the bounding box of previously placed word j:

$$\chi_{B_j}(x, y) = \begin{cases} 1 & \text{if } (x, y) \in B_j \\ 0 & \text{otherwise} \end{cases}$$

The final plots for each time point were combined with the generated embedded words to produce comprehensive visual representations of the trends and keywords in the dataset. These visualizations were saved as images for further analysis and presentation.

2.6 Data Analysis and Evaluation

The algorithm was implemented in Python. After generating visualizations, we compared our results with two conventional techniques, then we conducted a case study with data analysis for demonstration

Interpretation of Plots: The interpolated curves ($P_{smooth}(x)$) plotted for each method provided a visual representation of the document trends over time. These curves were accompanied by shaded areas ($P_{lower}(x)$ and $P_{upper}(x)$), denoting the uncertainty or variability in documents occurrences.

Analysis of embedded words: Word embedding elements generated from keywords associated with each year-category combination ($K(y,c)$) provided insights into the prevalent research topics and themes. We identified *Keyword Frequency* in which prominent keywords indicated the frequency of topics addressed within each category over time.

3 Results and Discussion

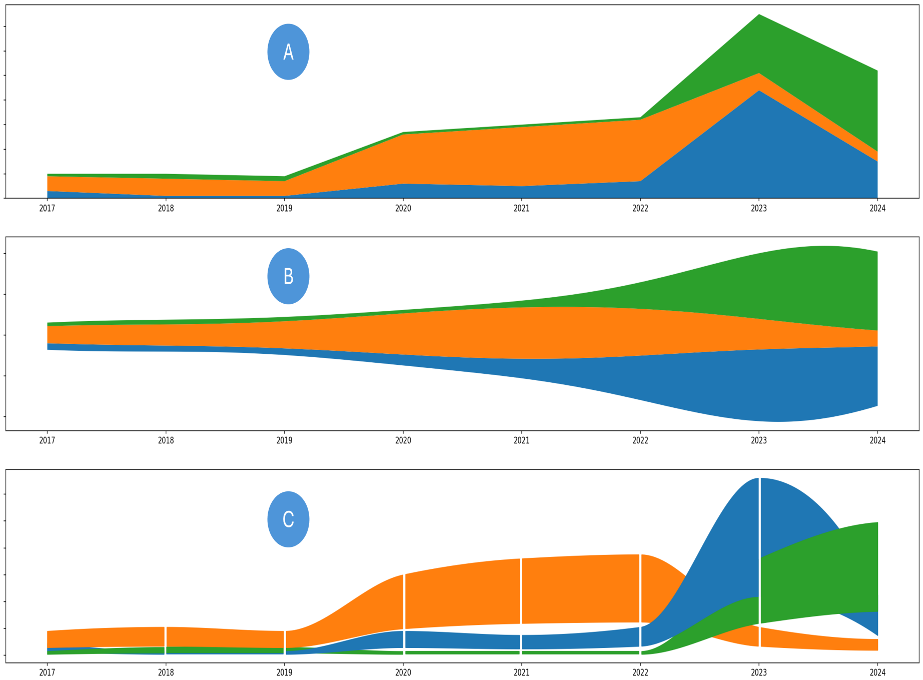

Fig. 2. (A) Stacked Area Chart, (B) Stream Graph, (C) Our work: Sorted Stream Graph on the same data

Fig. 2 presents an intricate comparison between traditional data visualization methodologies (A and B) and our refined technique (C). The early years (2017-2022) of our study indicate that users faced minimal challenges when deciphering data visually. Dominance by the orange color within these graphs is attributed to its prominent data variation. However, as we advance into 2023 and beyond, notable distinctions between categories become apparent, particularly when examining blue and green on stacked area chart and stream graphs. Our methodological enhancement of organizing the graph segments in ascending order simplifies comparative analysis – with greater values positioned above lesser ones within each segment, thereby enhancing readability. By 2024, even when both visualization tools (stacked area chart and stream graph) render blue and green data variation nearly indistinguishable, our refined sorted stream graph maintains its ability to differentiate between these similar fluctuations with greater clarity. Additionally, this method allows for the concurrent monitoring of topic trends – a capability that is absent in both stacked area charts and standard stream graphs.

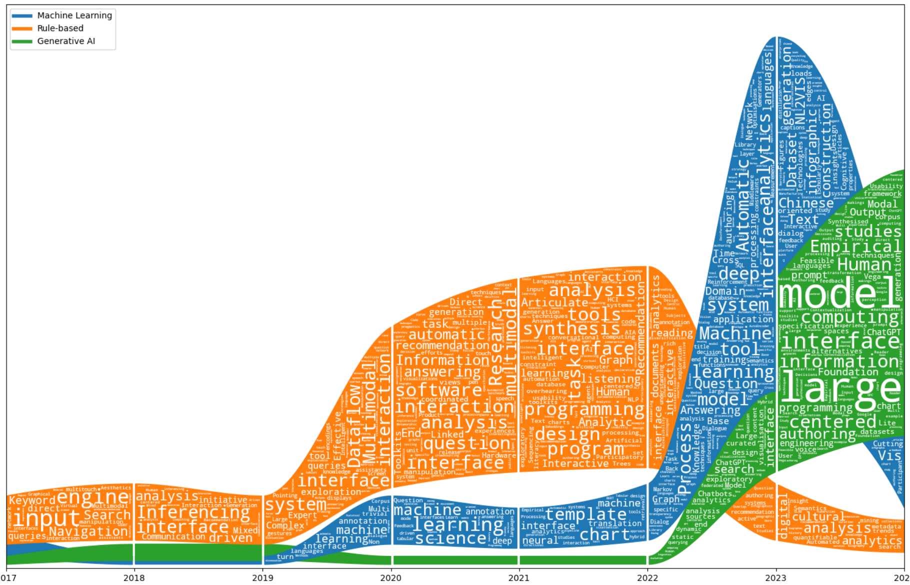

Fig. 3. The evolution of automatically generated data visualizations

Case study: The evolution of automatically generated data visualizations In this study, our interested research question is "*How has the progression of automatically generated data visualizations evolved over the years? taking into account of rule-based, machine learning and generative AI approach*". To answer this research question, we used Google Scholar as an indexing database for searching relevant publications. Starting from a paper pioneering in this topic

[5], we used Snowball technique to discover new relevant publications by looking all publications citing it. We did this manually by going through ALL citations until no citation from relevant paper is presented. As a result, there are 104 publications highly relevant to this topic.

Fig. 3 presents a clear depiction of our proposed approach's outcomes. From 2017 to 2022, the visualization landscape was dominated by rule-based methods, with machine learning techniques trailing behind. This dominance likely arose from the maturity and reliability of rule-based systems, which are often simpler to implement and understand compared to their machine learning counterparts. During this period, frequent terms related to rule-based methodologies such as *engine, input, interface, dataflow, interaction, multimodel, design, program, and analysis* were prominent, highlighting the emphasis on these core components within the visualization domain. However, interest in machine learning approaches began to rise, hinting at a transition towards more advanced and dynamic visualization techniques. This shift was fueled by the increasing availability of large datasets and enhanced computational power, which enabled more complex and responsive visual representations.

The period from 2022 to 2023 marked a significant turning point. Publications utilizing machine learning approaches surged, surpassing those based on rule-based methods. Generative AI followed closely, while rule-based learning publications dwindled. This dramatic change can be attributed to the rapid advancements and successes in machine learning and generative AI. Machine learning's ability to learn from vast amounts of data and adapt to new patterns offered significant advantages over traditional rule-based methods. This adaptability led to more accurate and insightful visualizations, capable of uncovering hidden trends and correlations. The rise of generative AI, with its ability to create new data and enhance existing datasets, further amplified the potential for novel visual representations. Keywords related to machine learning, such as *annotation, templates, chart, process, deep, model, and analysis*, became increasingly common, reflecting this paradigm shift. The decline in rule-based learning publications was also due to their limitations in handling the complexity and volume of modern datasets. As data became more intricate and multifaceted, the static nature of rule-based systems struggled to keep pace. Consequently, researchers and developers increasingly turned to machine learning and generative AI to tackle these challenges, leading to a shift in publication trends.

By 2024,generative AI took center stage, with a large number of publications exploring its utilization. Terms like *large, model, human, empirical study, computing, and authoring* dominated the discourse. Generative AI demonstrated unprecedented capabilities in creating high-quality, realistic, and innovative outputs across various domains. The collaborative nature of the research community, coupled with the open-source movement, played a crucial role in accelerating the development and dissemination of generative AI technologies. The availability of pre-trained models, extensive datasets, and robust frameworks enabled researchers to experiment, iterate, and innovate at an unprecedented pace, contributing to the exponential growth in publications.

4 Conclusion

In this study, we introduced a novel visualization technique, the Sorted Stream Graph with Embedded Word Elements (SSGEW), designed to enhance the analysis of temporal data across multiple categories. It incorporates word embedding elements within the stream graph, providing a visual representation of data trends over time and integrating semantic insights for a more comprehensive understanding of underlying patterns. The SSGEW methodology addresses limitations of traditional stream graphs and stacked area charts, such as difficulty in cross-category comparisons and limited readability of overlapping segments. The sorted arrangement of categories within the stream graph facilitates easier comparative analysis, while the embedded word elements enrich the visualization with contextual information, aiding in the interpretation of trends and fluctuations. The SSGEW technique offers several advantages, including improved readability, enhanced comparative analysis, and the integration of semantic data. However, it has potential limitations, such as cross-overlaid graph, increased complexity of visualizations and the need for careful interpretation of embedded word elements. Future work may focus on optimizing overlays in graphs such as using small multiples (or subplots), or using interactive plots where users can hover over or click on data points.

References

1. Abi Haidar, A., Yang, B., Ganascia, J.G.: Visualizing the first world war using streamgraphs and information extraction. In: 2016 20th International Conference Information Visualisation (IV). pp. 290–293. IEEE (2016)
2. Anh, N.T.H., et al.: Access to online business opportunities: Enhancing digital technology capacity for women with disabilities in the red river delta of vietnam. Heliyon (2024)
3. Bostock, M., Ogievetsky, V., Heer, J.: D^3 data-driven documents. IEEE transactions on visualization and computer graphics **17**(12), 2301–2309 (2011)
4. Byron, L., Wattenberg, M.: Stacked graphs–geometry & aesthetics. IEEE transactions on visualization and computer graphics **14**(6), 1245–1252 (2008)
5. Cox, K., Grinter, R.E., Hibino, S.L., Jagadeesan, L.J., Mantilla, D.: A multi-modal natural language interface to an information visualization environment. International Journal of Speech Technology **4**, 297–314 (2001)
6. Dang, T., Nguyen, H.N., Pham, V., Johansson, J., Sadlo, F., Marai, G.: Wordstream: Interactive visualization for topic evolution. In: EuroVis (Short Papers). pp. 103–107 (2019)
7. Deng, Q., Cai, G., Zhang, H., Liu, Y., Huang, L., Sun, F.: Enhancing situation awareness of public safety events by visualizing topic evolution using social media. In: Proceedings of the 19th annual international conference on digital government research: Governance in the data age. pp. 1–10 (2018)
8. Fritsch, F.N., Carlson, R.E.: Monotone piecewise cubic interpolation. SIAM Journal on Numerical Analysis **17**(2), 238–246 (1980)
9. Munzner, T.: Visualization analysis and design. CRC press (2014)

10. Nguyen, V.T., Jung, K., Gupta, V.: Examining data visualization pitfalls in scientific publications. Visual Computing for Industry, Biomedicine, and Art **4**, 1–15 (2021)
11. Shneiderman, B.: The eyes have it: A task by data type taxonomy for information visualizations. In: The craft of information visualization, pp. 364–371. Elsevier (2003)
12. Tufte, E.R.: The visual display of quantitative information. The Journal for Healthcare Quality (JHQ) **7**(3), 15 (1985)
13. Wickham, H.: ggplot2. Wiley interdisciplinary reviews: computational statistics **3**(2), 180–185 (2011)