A Heatmap-Based Time-Varying Multi-Variate Data Visualization Unifying Numeric and Categorical Variables

Haruka Suematsu, Sayaka Yagi, Takayuki Itoh
Ochanomizu University
Tokyo, Japan
{*haruka, sayaka, itot*}*@itolab.is.ocha.ac.jp*

Yosuke Motohashi, Kenji Aoki, Satoshi Morinaga
NEC Corporation
Tokyo, Japan
y-motohashi@bk.jp.nec.com, k-aoki@bq.jp.nec.com, morinaga@cw.jp.nec.com

***Abstract*—Most time-varying data in our daily life is multi-variate. Moreover, most of such time-varying data contains both numeric and categorical values. It is often meaningful to visualize both of them as they are often correlated. We aim to visualize every value in such time-varying data in a single display space so that we can discover interesting relationships among the values of the time-varying data. This paper presents a heatmap-based time-varying data visualization technique which displays both numeric and categorical values in a single display space. The technique assigns time to the horizontal axis of the display space, and vertically arranges the series of colored belts corresponding to the time-sequence values. It generates one belt for a numeric value, and multiple belts for a categorical value. It clusters the belts according to the similarity of color sequences, and re-arranges the belts based on the clustering result. This paper shows an example of the visualization result applying a time-varying multi-variate marketing dataset.**

***Keywords*-Visualization; Heatmap; Time-varying data; Multi-variate data; Clustering**

I. Introduction

Recent evolutions of computer simulation techniques, sensing technologies, and large-scale data storage technologies have brought an increase in huge and complex time-varying data. It is not always easy to effectively discover a fruitful knowledge even applying the latest data mining techniques. Time-varying data visualization is useful to subjectively explore such data and discover a fruitful knowledge. Polyline and heatmap are the most popular approaches to represent the time-varying data. This paper focuses on heatmap-based time-varying data visualization.

Most time-varying data in our daily life is multi-variate. For example, weather measurement data can be treated as time-varying data containing multiple values, such as weather (e.g. fine, cloudy, rainy), temperature, humidity, and amount of rainfall. Moreover, many of such time-varying data contains both numeric and categorical values. In the above example, weather is categorical while temperature and humidity are numeric. It is often meaningful to visualize both of them as they are often correlated. We aim to visualize every value in such time-varying data in a single display space so that we can discover interesting relationships among the values of the time-varying data.

This paper presents a heatmap-based time-varying data visualization technique which displays both numeric and categorical values in a single display space. The technique assigns time to the horizontal axis of the display space, and vertically arranges the series of colored belts corresponding to the time-sequence values. It generates one belt for a numeric value, and multiple belts for a categorical value. It clusters the belts according to the similarity of color sequences, and re-arranges the belts based on the clustering result. This design is useful to visually recognize the correlations among the multiple values in the time-varying datasets.

II. Related Work

A. Multi-variate Data Visualization

There have been variety of multi-variate visualization techniques. Non-time-varying multi-variate datasets can be effectively visualized in a single display space, by applying Parallel Coordinates or Scatterplot-Matrices. Dimension analysis techniques have been applied to these visualization techniques, to determine meaningful orders of dimensions for Parallel Coordinates, or to select meaningful pairs of dimensions for Scatter Plots.

Parallel Coordinates or Scatterplot can be extended as 3D time-varying multi-variate data visualization systems by assigning time to the third axis [1]. However, these techniques have an essential bottleneck that 3D visualization techniques often cause cluttering problems.

B. Time-varying Data Visualization

Polyline is the most common representation for time-varying data in our daily life. However, a polyline-based representation has two drawbacks:

1) cluttering among large number of polylines in a single display space, and
2) utilizing less display space when the numeric distribution is unbalanced.

While several polyline-based techniques [9] are improved to solve the above problems, other representations including heatmaps have been applied to many time-varying data

1550-6037/14 $31.00 © 2014 IEEE
DOI 10.1109/IV.2014.25

visualization techniques. Also, clustering of time-varying data is also effective to sample large-scale datasets.

Other representations for time-varying data visualization include 3D histogram [6], spiral representation for periodicity analysis [8], piles of painted polylines such as ThemeRiver [3], and Two-Tone Pseudo coloring [7].

Heatmap-based representation has been also applied to various time-varying data visualization techniques, as well as the technique presented in this paper. Heatmap has advantages over other representations from the standpoint of cluttering reduction and display space utilization for overviews. Imoto et al. presented a technique that extracts interesting portions of time-varying data on a heatmap [5]. Ziegler et al. [10] also presented a heatmap-based technique applying Pixel Bar Charts. These techniques do not focus on multi-variate datasets containing both numeric and categorical variables.

C. Visual Analytics Tools for Time-Varying Data

There have been several studies on development of heatmap-based visual analytics tools for time-varying data. WireVis [2] is a coordinated-view system of heatmap, polyline charts, pie charts, and search results, which are applied to visualize transaction data. Hayashi et al.[4] presented a similar system featuring a variable recommendation algorithm so that users can easily find interesting trends in particular variables. However, these systems do not simultaneously display multiple attributes of time-varying multivariate datasets. Thus, users need to determine the attributes and select one of them to be displayed in advance, and therefore, it is difficult to analyze relevancy among multiple attributes from the visualization results.

The technique presented in this paper displays such multiple values considering of mixture of numeric and categorical variables, so that it allows users to simultaneously compare the time sequence of multiple attributes.

III. Presented Visualization Technique

This section describes the overview and processing flow of the presented visualization technique.

A. Data Structure

We aim to visualize time-varying data which contains both numeric and categorical values. We suppose a set of events $E = \{e_1, ..., e_{ne}\}$ is given as an input dataset, where E has n_e events. We describe an event as follows:

$$e_i = \{t, v_1, ..., v_{m_v}, c_1, ..., c_{m_c}\},$$

where t is the time when the event occurs, v_j is the j-th numeric value of the event, c_j is the j-th categorical value of the event, m_v is the number of numeric values, and m_c is the number of categorical values. We describe choices of a categorical value as $c_j = \{C_{j1}|C_{j2}|...|C_{jn_{cj}}\}$, where n_{cj} is the number of choices for the j-th categorical value.

Our implementation constructs a time-varying data from the above-mentioned collection of events. Here, we describe the timestamps $\{t_0, t_1, ...\}$. The implementation collects the events occurred during the i-th time span $[t_i, t_{i+1}]$. It calculates the average values for each numeric variable, and frequency for each categorical variable, from the collected events. This section describes the average of the j-th numeric value in the i-th time span as $\overline{v}_{ji}$, and the frequency of C_{jk} in the i-th time span as $\overline{C}_{jki}$. Our implementation generates time sequences of

$$\text{numeric values } \boldsymbol{V}_j = \{\overline{v}_{j1}, \overline{v}_{j2}, ...\} \text{ and}$$
$$\text{categorical values } \boldsymbol{C}_{jk} = \{\overline{C}_{jk1}, \overline{C}_{jk2}, ...\},$$

and then displays them as colored belts. Figure 1 denotes the data structure.

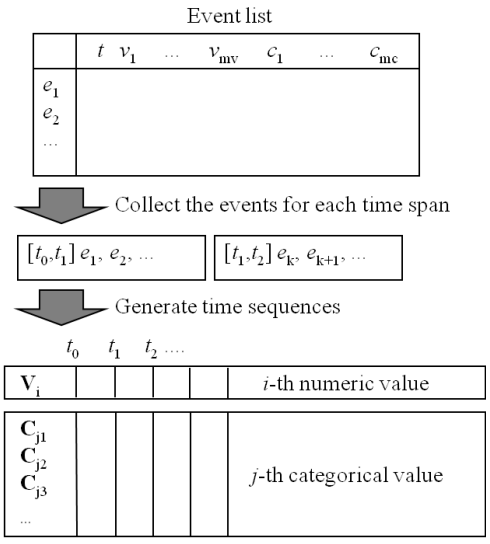

Figure 1. Time-varying data construction from an event list.

B. Overview

Figure 2 shows a snapshot of the presented technique. The average of numeric values or frequency of categorical values is represented as colored thin belts. In other words, the j-th numeric value $\boldsymbol{V}_j$ or the k-th choice of the j-th categorical value $\boldsymbol{C}_{jk}$ corresponds to a thin belt in the visualization result. The horizontal axis denotes the timeline, while the belts are arranged along the vertical axis based on their similarities.

The left side of the window features GUI widgets so that users can freely customize the visualization results. It can control the following: zooming operations, color mapping rule, choices of time span (e.g. hour, day, day of the week, month), and visible/invisible mode selection for each of the numeric or categorical values.

C. Heatmap Coloring and Rearrangement

As a result of the data conversion shown in Figure 1, we suppose each of the following variables,

$$\{\boldsymbol{V}_1, ..., \boldsymbol{V}_{mv},$$
$$\{\boldsymbol{C}_{11}, \boldsymbol{C}_{12}, ...\}, \{\boldsymbol{C}_{21}, \boldsymbol{C}_{22}, ...\}, ..., \{\boldsymbol{C}_{mc1}, \boldsymbol{C}_{mc2}, ...\}\}$$

are represented as colored thin belts. Our technique assign colors to the each variable, and rearranges the order of the belts based on their similarity.

Our current implementation firstly normalizes the values. Numeric values at each span $\{\overline{v}_{j1}, \overline{v}_{j2}, ...\}$ are normalized to $[0, 1]$ using their minimum and maximum values. Categorical values at each span $\{\overline{C}_{jk1}, \overline{C}_{jk2}, ...\}$ are also normalized to $[0, 1]$ using their minimum and maximum frequency values. Our implementation then assigns colors from the normalized values by using the same color map. Here, the normalized numeric or categorical values can be treated as n-dimensional vectors, if there are n time spans. Our implementation divides the vectors into adequate number of clusters based on their similarities, and closely displays colored belts corresponding to the vectors belonging to the same cluster.

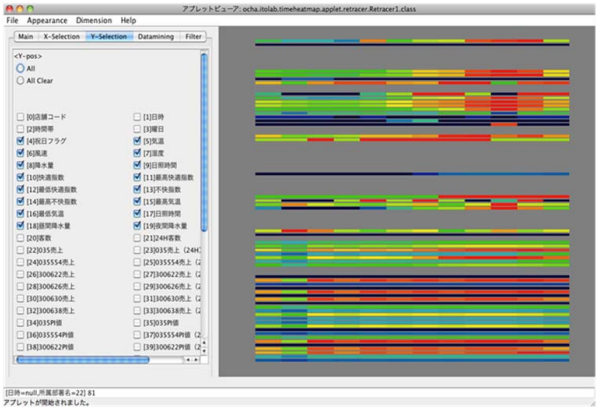

Figure 2. Snapshot of the presented technique.

IV. Example

We implemented the technique with Java Development Kit (JDK) 1.7.0 and Java binding OpenGL (JOGL) 2.0. We executed the implementation on Apple MacBookPro with Max OS 10.6.

This section introduces an example of visualization applying to a transaction dataset. The dataset contained 5440 events consisting of 57 numeric values and 11 categorical values. Numeric values included air temperature, humidity, and the number of sold products. Categorical values included types and names of product, shops, and locations.

Figure 3(Left) shows the heatmap before applying the clustering of the colored belts. Figure 3(Center) shows the heatmap after dividing the variables into 14 clusters. This result closely displays similarly looking belts so that we can easily compare the correlated variables. Figure 3(Right) shows the heatmap after interactively selecting the variables to be displayed, and dividing the displayed variables into 9 clusters. Annotations in blue drawn in the figure denote the clusters. This result demonstrates the technique adequately displays multiple sets of correlated time-varying values.

Figure 4 shows a close up view to the particular values. Our dataset contained a categorical value "name of product". We observed the particular cluster containing the numeric variables "number of customers", "air temperature", "humidity", and "discomfort index", shown as the upper four belts in Figure 4. Besides the category "A" of the categorical value, "name of product" belonged to the cluster. Figure 4 denotes that actually these values look similar. On the other hand, other categories "B", "C", "D", "E", and "F" of "name of product" did not belong to the above mentioned cluster, and actually their values did not look similar to the values of the variables in the cluster. We found only the sale of product "A" correlates to the number of customers and weather conditions. The technique brought such knowledge by decomposing the time-varying categorical values into each category, and applying the clustering process.

V. Conclusion

This paper presented a heatmap-based visualization technique for time-varying multi-variate datasets, which contains both numeric and categorical values. The technique generates colored belts for each numeric value or for each category of the categorical values, and displays all the belts in the order of their clustering result. We would like to have more experiments with variety of applications, test the scalability, and conduct subjective evaluations of the technique.

References

[1] E. W. Bethel, O. Ruebel and G. Weber, *Visualization and Analysis of 3D Gene Expression Data*, Two-page research highlight, Lawrence Berkeley National Laboratory, LBNL-63658, 2007.

[2] R. Chang, M. Ghoniem, R. Kosara, W. Ribarsky, J. Yang, E. Suma, C. Ziemkiewicz, D. Keim and A. Sudjianto, *WireVis: Visualization of Categorical, Time-Varying Data from Financial Transactions*, IEEE Symposium on Visual Analytic Science and Technology, 155-162, 2007.

[3] S. Havre, B. Hetzler and L. Nowell, *ThemeRiver: Visualizing Theme Changes over Time*, IEEE Symposium on Information Visualization, 115-123, 2000.

[4] A. Hayashi, T. Itoh and S. Nakamura, *A Visual Analytics Tool for System Logs Adopting Variable Recommendation and Feature-Based Filtering*, 17th International Conference on Information Visualisation (IV2013), 1-10, 2013.

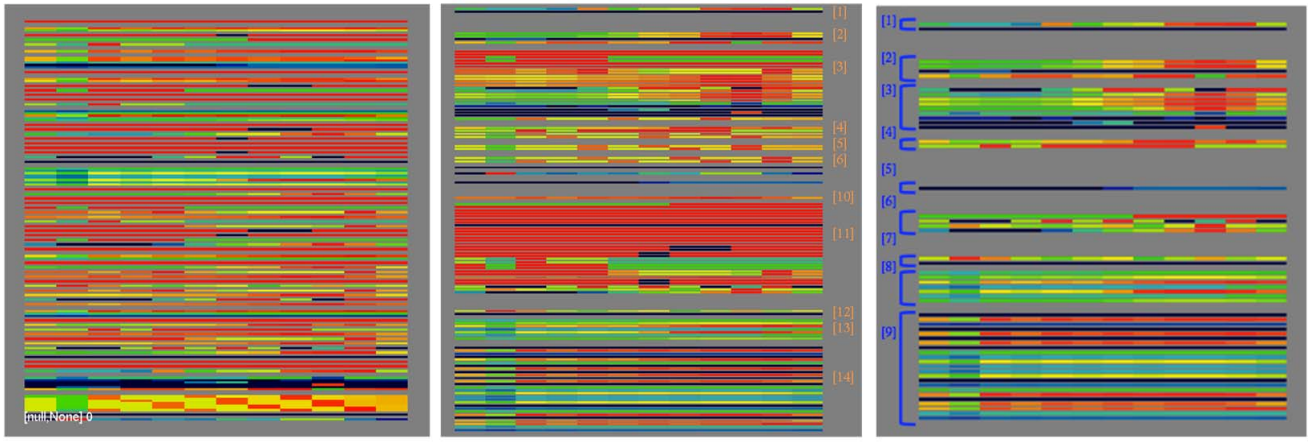

Figure 3. Example of time-varying data visualization constructed from a transaction dataset. (Left) before clustering variables. (Center) After clustering variables. (Right) After manually selecting interested variables.

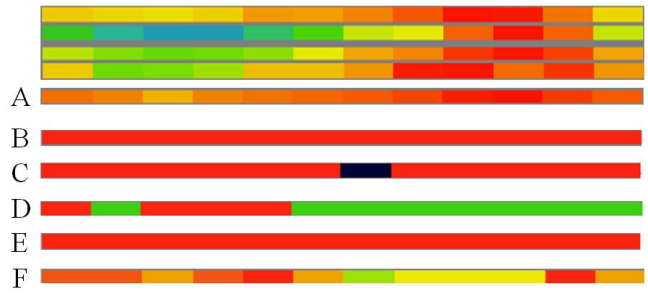

Figure 4. Close up view to the particular time-varying values. A Categorical value contains ”A”, ”B”, ”C”, ”D”, ”E” and ”F”, and only ”A” belongs to the cluster with the upper four time-varying value. We found only the category ”A” correlates to the upper four values by observing our visualization result.

[5] M. Imoto and T. Itoh, *A 3D Visualization Technique for Large Scale Time-Varying Data*, 14th International Conference on Information Visualisation (IV2010), 17-22, 2010.

[6] R. Kosara, F. Bendix and H. Hauser, *Timehistograms for Large, Time-Dependent Data*, Eurographics/IEEE TVCG Symposium on Visualization, 45-54, 2004.

[7] T. Saito, H. Miyamura, M. Yamamoto, H. Saito, Y. Hoshiya and T. Kaseda, *Two-Tone Pseudo Coloring: Compact Visualization for One-Dimensional Data*, IEEE Symposium on Information Visualization, 173-180, 2005.

[8] M. Weber, M. Alexa and W. Muller, *Visualizing Time-Series on Spirals*, IEEE Symposium on Information Visualization 2001, 7-14, 2001.

[9] S. Yagi, Y. Uchida and T. Itoh, *A Polyline-Based Visualization Technique for Tagged Time-Varying Data*, 16th International Conference on Information Visualisation (IV2012), 106-111, 2012.

[10] H. Ziegler, M. Jenny, T. Gruse and D. A. Keim, *Visual Market Sector Analysis for Financial Time Series Data*, IEEE Symposium on Visual Analytics Science and Technology, 83-90, 2010.