Multi-Resolution Techniques for Visual Exploration of Large Time-Series Data

Ming Hao and Umeshwar Dayal
HP Laboratories, Palo Alto, CA

Daniel Keim and Tobias Schreck
University of Konstanz, Germany

1 MOTIVATION

Time-series are a data type of utmost importance. In many domains like business applications, process monitoring, engineering, and security, huge amounts of complex, time-related data are collected and analyzed. Automatic algorithms are one option for dealing with time-series data, e.g., by searching for predefined patterns in time-series, or by fitting statistical models to the data for trend analysis and prediction. Often, it is not clear what the relevant patterns to look for are. Rather, the analyst needs to perform assumption-free exploration of the data to search for interesting patterns and to find previously unknown information.

The standard approach using bar- or line-charts without any data preprocessing is ineffective for visual analysis of large time-series databases: Given the limits of current display devices we either have to accept overplotting (occlusion) effects in the display or we have to aggregate the data, which may lose important information. For example, in Figure 1 (top), we show a time-series of 7,701 values. Without data reduction, we observe severe overplotting effects. With data reduction, e.g., by equally spaced sampling or averaging, information loss is introduced, but the lost information may be very important, as in network monitoring where the peaks (extreme) are of interest.

Previous work on visualization of long time-series has focused on numeric aggregation and on devising new space-efficient rendering methods. Common to these approaches is that they apply a uniform resolution level in aggregating and drawing. They do not allow for locally varying degrees of aggregation.

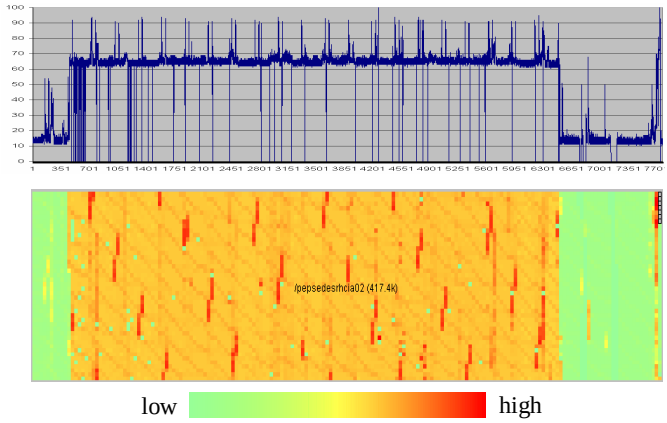


Figure 1: Visualize 28 days worth of network monitor probes from one server resulting in 7,701 observation values. On the top, the raw data shows overplotting. On the bottom, each observation is represented by a color-cell. The color of a cell represents the value of an observation. Cells are arranged from bottom to top (column) and left to right. This color-cell time-series visualization is able to display each observation without losing information.

2 OUR APPROACH

In this work, we introduce an importance-driven distortion technique to generate multi-resolution layouts for long time-series data. First, we employ a color-coded layout shown in Figure 1 (bottom). The color-coded cell-based matrix technique displays each data value by a rectangular amount of screen space. The rectangle (cell) is color-coded to represent the magnitude of the value to be shown. The display in Figure 1 (bottom) is able to visualize the full data set without data reduction, thereby not incurring any information loss.

Second, we apply the notion of the *intra time-series degree of importance* (DOI) to generate multi-resolution layouts for long time-series data as shown in Figure 2. The proposed layouts display data intervals considered important at high resolution levels with respect to the visualization method, enabling the analyst to quickly perceive important data characteristics in the time-series. At the same time, the layout displays the remaining, comparatively less important data intervals at lower resolution, providing the context of the full time-series data.

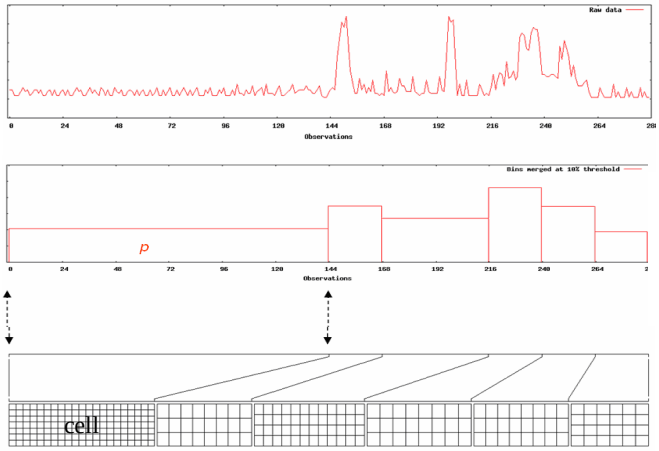


Figure 2: Map the degree of importance of a server process utilization time-series to a corresponding multi-resolution layout. This technique divides data into different time intervals. Then, these intervals are mapped into the screen partitions. Each partition consists of a number of cells representing data from that interval. The dimension of a partition reflects the level of resolution. Small cells are used for low resolution levels.

We have applied this technique to two different types of applications: time-driven (Section 2.1) and data-driven (Section 2.2) usage cases, based on the aspect from which the time-series importance is derived.

2.1 Time-Driven Multi-Resolution Layout & Applications

In the time-driven case, we specify the importance as a function of data age. We generate corresponding multi-resolution layouts where the level of resolution for each data partition to be

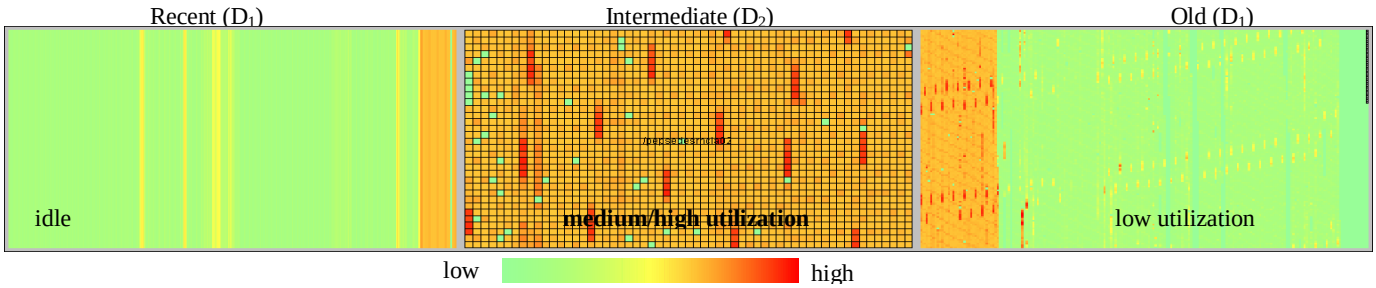


Figure 3: Show an Internet server utilization time-series.

visualized depends on the corresponding data age. We define the *distorted time-series importance profile (DOI profile)* as a set of triples called multi-resolution index *MRI*, each indicating an amount of data D_i, an amount of display space R_i, and a rendering method V_i for visualizing a data interval:

$MRI_i = (D_i, R_i, V_i)$ $\quad i=[1,...,n]$ $\quad n$: number of data partitions

We apply the time-dependent multi-resolution technique on a large real-world data set from the network monitoring domain as illustrated in Figure 3. For this application, we define a three-fold multi-resolution profile using relative weights for data and display:

$$MRI = \{ MRI_1, MRI_2, MRI_3 \}$$
$$= \{(1, 1, \text{c-matrix}), (4, 1, \text{c-matrix}), (16, 1, \text{c-matrix})\}$$

This means that the whole time-series is divided into three intervals. Each interval is given the same amount of display space, as $R_i=1$ for all i. The partitions contain increasingly more data, such as: $D_1=1$, $D_2=4$, $D_3=16$. The display emphasizes perception of the most recent data interval D_1 at high resolution (data-to-display relation is *1:1*), while maintaining the intermediate (D_2) and old data (D_3) in context (data-to-display relations are *4:1* and *16:1*).

Figure 3 visualizes *25,920* process utilization values taken from one target Internet server on a network. The display allows visually analyzing large time intervals, keeping the most recent data (leftmost partition) at the highest resolution. The cell sizes for the older data are automatically scaled down via increasing data density. The color of a cell represents the value of the server utilization in each observation from low to high (green, yellow, orange, and red). From Figure 3, we learn that the server currently is idle (green), but before that the server experienced medium/high utilization (yellow, orange, and red) over one third of the time. During that time, hot spots (red) occurred in a repeated regular pattern in D_2.

2.2 Data-Driven Multi-Resolution Layout & Application

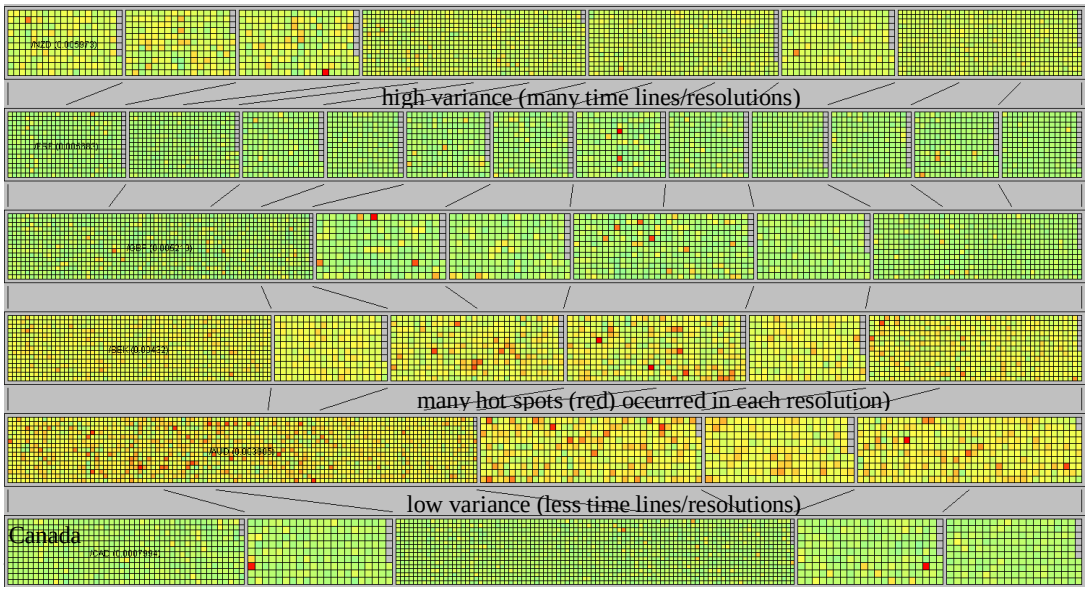
Often, there exist scenarios where the user has no a-priori knowledge regarding the data, but expects the visualization to structure (guide) the data exploration and analysis process. To this end, we propose to have an algorithm automatically analyze the data and derive an appropriate data-dependent DOI profile on the fly. From our experiments, we found that the number of resolution partitions should not be too high. Our technique enables the user to quickly identify (focus on) a few, but top important portions of the data, while keeping the complete time-series in perspective.

Figure 4 shows the time- series of exchange rates of six International currencies measured against the US-$ for roughly 10 years, corresponding to 2,566 exchange rates per currency [1]. We set the profile generator to consider the *maximum* aggregation function for generating variable time slice layouts independently for each series, using *time lines* as visual clues aligning the time intervals. Our technique isolates time subintervals with sufficiently different maximum exchange rates for the different currencies. We recognize that the Canadian $, as given in the last row in Figure 4, has the most homogenous exchange rates with little variance and is segmented into only 5 different subintervals. The other five currencies show more dynamics in the exchange rates, as is obvious from the many different time slices.

3 Conclusions

In this poster, we have introduced an importance-driven distorted partition technique for visualizing large time-series data. Our technique employs non-linear rescaling in conjunction with a space-efficient new rendering method. Rescaling was performed by generating either time-dependent or data-dependent layouts based on data importance. Our future work will incorporate clustering methods for ordering multiple time-series layouts.

[1] Spot Exchange Rate Dataset. Institute of Statistics and Decision Sciences, Duke University, 2005, Durham, NC.



- Each cell represents a time interval.
- Cell size in each partition represents the resolution.
- Color represents the exchange rate.
- Each partition in a time-series is connected by a *time line*

Figure 4: Show variable time slice results for the exchange rate of six International currencies measured against US-$ using the MAXIMUM aggregation function and independent DOI analysis for each of the time-series.