Visual Analysis of Frequent Patterns In Large Time Series

M. C. Hao[1], M. Marwah[1], H. Janetzko[2], D. A. Keim[2], U. Dayal[1], R. Sharma[1], D. Patnaik[3], N. Ramakrishnan[3]
[1]Hewlett Packard Laboratories, USA [2]University of Konstanz, Germany [3]Virginia Tech, USA

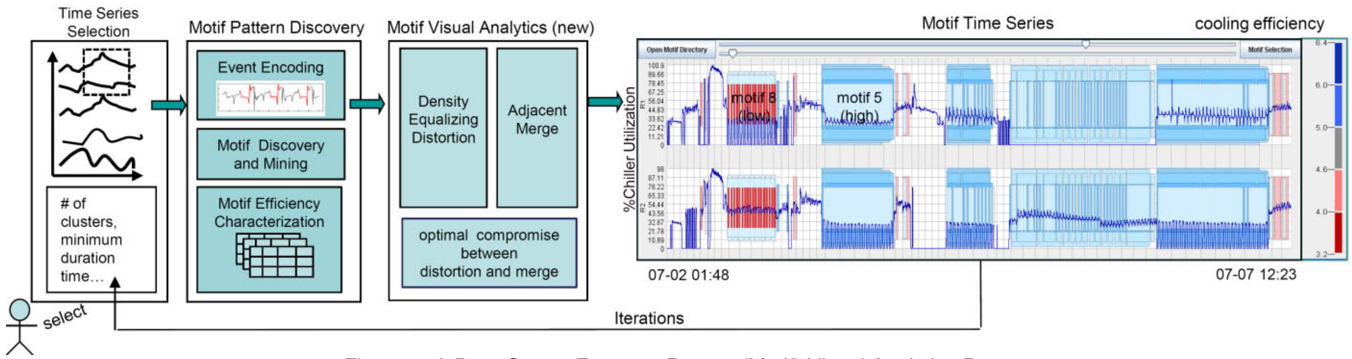


Figure 1: A Data Center Frequent Pattern (Motif) Visual Analytics Process
(x-axis: time in 1-minute intervals, y-axis: %utilization of chillers R1 and R2, color: chiller's cooling efficiency from low (red) to high (blue))
Motifs are represented by rectangles. The height of a motif is proportional to the average duration of all occurrences of the same motif.

ABSTRACT

The detection of previously unknown, frequently occurring patterns in time series, often called motifs, has been recognized as an important task. To find these motifs, we use an advanced temporal data mining algorithm. Since our algorithm usually finds hundreds of motifs, we need to analyze and access the discovered motifs. For this purpose, we introduce three novel visual analytics methods: (1) **motif layout**, using colored rectangles for visualizing the occurrences and hierarchical relationships of motifs in a multivariate time series, (2) **motif distortion,** for enlarging or shrinking motifs as appropriate for easy analysis and (3) **motif merging,** to combine a number of identical adjacent motif instances without cluttering the display. We have applied and evaluated our methods using two real-world data sets: data center cooling and oil well production.

1. Introduction

Efficient algorithms for detecting motifs in time series data have been used in many applications, such as detecting anomalies in patients' medical records over time. To visualize motifs, Lin's VisTree [1] transforms a large time series into a symbolic representation, then encoding the data into a tree with branches to represent symbols and motifs. However, analysts want to have an overview of motifs within a single view and search for the most efficient motif based on a performance metric.

———————————————

IEEE Symposium on Visual Analytics Science and Technology
October 24 - 29, Salt Lake City, Utah, USA
978-1-4244-9487-3/10/$26.00 ©2010 IEEE

Figure1 illustrates process can be subdivided into three phases: (1) the input selection phase to select the multivariate time series and parameters, (2) the motif pattern discovery phase to map a multivariate time series to frequent patterns (motifs), such as the motif starting/ending time and its efficiency value, (3) the motif visual analytics phase to layout the discovered motifs back into a multivariate time series.
In Figure 1, motifs are represented by rectangles of different sizes. They can be of varying lengths, with many shorter motifs nested within longer motifs. Each motif is specified in terms of the starting and ending times of the pattern. We allow users to adjust the degree of distortion and merge to generate the best view for analyzing the time series. In addition, we link the motifs to the associated cooling efficiency metrics for administrators to detect the most or least efficient motifs (e.g., motif 5 and 8) for further analysis.

2. Motif Pattern Discovery

Event encoding: We perform a k-means clustering on the multivariate time series considering each time point as a vector and use the cluster labels as symbols to encode the time series [2].

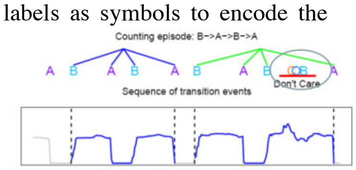


Figure 2: Example of Non-Overlapping Counting

Motif mining: *Frequent* episode mining is conducted on the transition event stream to detect repetitive motifs. Figure 2 gives an example of two non-overlapped occurrences of the motif B→A→B→A.

Efficiency Characterization: Each motif is characterized in terms of user defined efficiency metric and colors, which quantifies the heat removed per unit energy consumption. This enables efficiency comparisons between motifs to categorize them as good or bad.

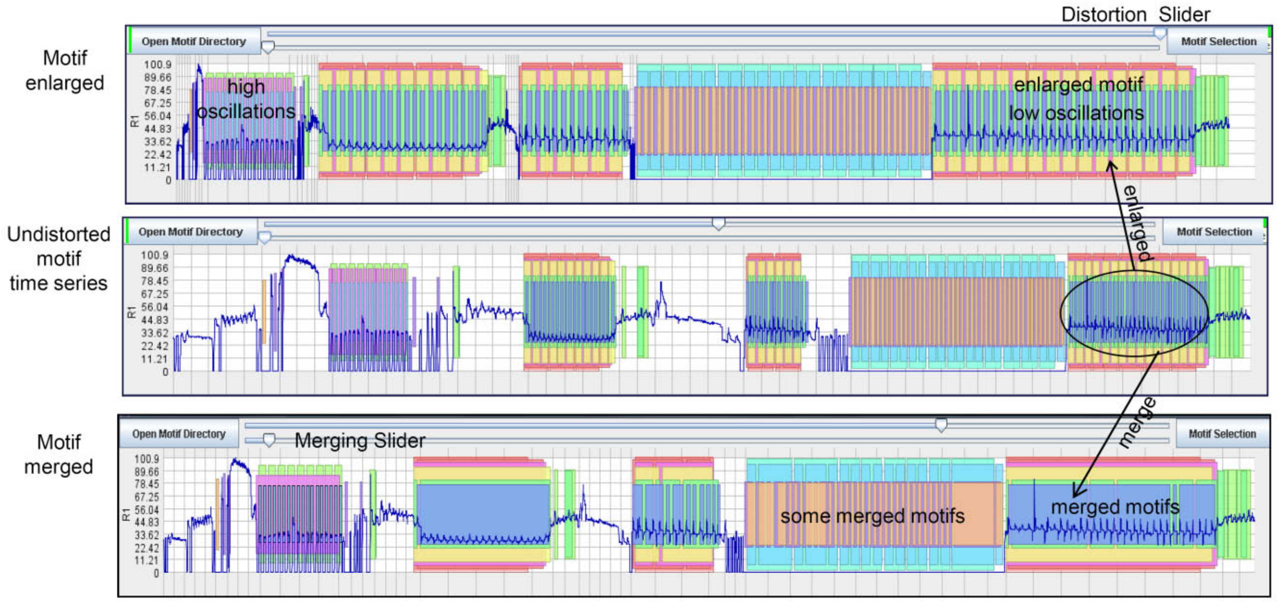


Figure 3: Motif Visual Distortion and Merging
(x-axis: time intervals, y-axis: %utilization of chiller R1, rectangles: motifs, color: motif types)
If distortion slider is moved to the right, motifs are enlarged. If merging slider is moved to the right, adjacent motifs are merged.

3. Motif Visual Analytics

Layout: To visualize motifs in a large complex time series, we derive a new layout algorithm and draw rectangles to represent the occurrences of motifs. The color of a rectangle represents its importance (e.g., efficiency). The nested rectangles are used for visualizing the hierarchical relationships among motifs. The height of a rectangle is linearly proportional to the statistical rank of the average duration time of the motifs. For users to easily analyze large number of nested motifs, we introduce two user interactions:

Distortion: Distortion enlarges the selected motifs using a user-activated distortion slider shown in the top of motif of Figure 3. Distorting the time series is done by applying an adapted density-equalizing distortion technique. We calculate weights for each time interval and use them as the input to the distortion algorithm. When the user moves the distortion slider to the right, the areas with motifs are enlarged.

Merging: We also provide a second slider in the bottom motif of Figure 3 to merge multiple motif occurrences to a single rectangle to reduce the visual clutter. If the slider is moved, motifs of the same type that begin or end at adjacent positions are combined.

After applying various degrees of distortion and merging, the motif time series is greatly simplified for visual analytics.

4. Applications and Evaluation

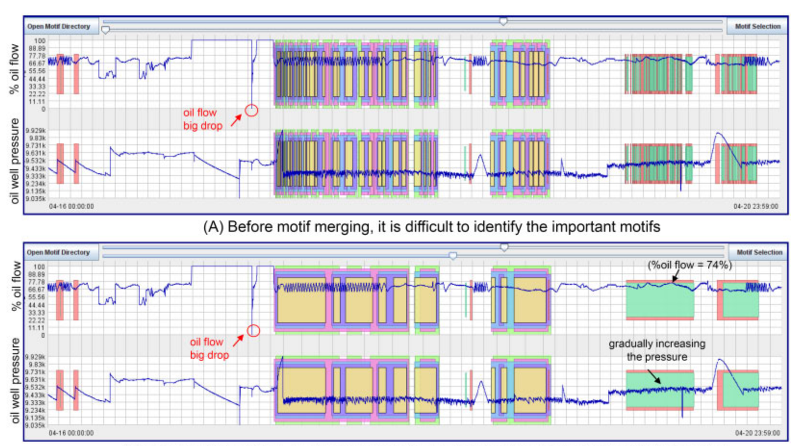
We have applied above techniques to analyze data center chiller sensor time series and oil well production sensor data. Figure 3 provides an enlarged view of the motifs, allowing data center administrators to compare them in terms of their oscillatory behavior which may decrease the life span of the chillers. From Figure 4B, oil-well production managers can easily see that the green motif is the most productive with an oil flow of up to 74%. Also, the production manager can determine that after a big drop in oil flow, it is best to gradually increase the pressure as shown in the green motifs. Inferring this from the raw time series can take days while using these motifs, it can be done in minutes.

5. Conclusion

Our results from both the real-world data center and oil/gas production time series sensor data show that our techniques successfully enable users to identify both efficient and inefficient motifs, and hence avoid inefficiencies.

[1] J. Lin et.al. VizTree: A Tool for Visually Mining and Monitoring Massive Time Series Database. Proceedings of the 30th VLDB Conference, Canada, p#.1260, 2004.

[2] Patnaik, D., Marwah, M. Sharma, R. Ramakrishnan, N. Sustainable Operation and Management of Data Center Chillers using Temporal Data Mining. Proceedings of KDD'09, France.



(A) Before motif merging, it is difficult to identify the important motifs

(B) After motif merging, users are able to focus on import motifs to analyze their behavior.

Figure 4: Oil Well Production Time Series (85,035 records) with Seven Different Motifs
(x-axis: time, y-axis: % oil flow and pressure, color: motif type)