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

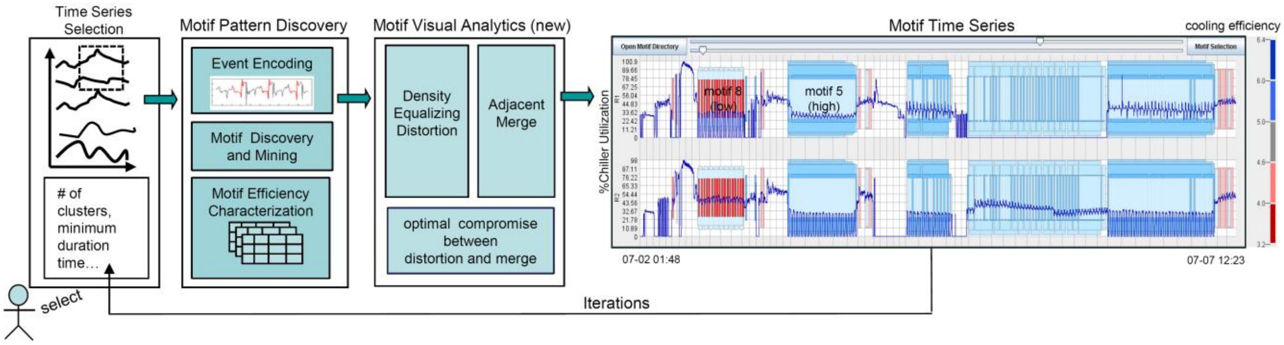


Figure 1: A Data Center Frequent Pattern (Motif) Visual Analytics Process
(x-axis: time in 1-minute intervals, y-axis: %utilization of chillers R1 and R2, color: chiller's cooling efficiency from low (red) to high (blue))
Motifs are represented by rectangles of a different size. The height of a motif 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 identifying words in different languages or detecting anomalies in patients' medical records over time. Figure 1 shows an example on how to monitor chiller efficiency in data centers using a pair of data center chiller time series in which different motifs were discovered. The illustrated 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. With our new motif distortion and merging techniques users are able to visualize their relationships and efficiency.

In the 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: Given the multivariate time series $T = \langle t_1,\ldots,t_m \rangle$ where each real valued vector t_i captures the data values (e.g., chiller utilization in the data center example), we first 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. Observe that the multivariate series is now encoded as a single (one-dimensional) symbol sequence [1].

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. The counting algorithm resets all internal states after completing each occurrence of the motif.

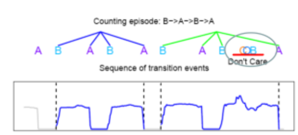


Figure 2: Example of Non-Overlapping Counting

Efficiency Characterization: Each motif is characterized in terms of the efficiency metric, which quantifies the heat removed per unit energy consumption. This enables efficiency comparisons between motifs: their categorization as 'good' or 'bad' from the efficiency metric point-of-view. Furthermore, this information helps to provide guidance to a manager regarding the most "efficient" configurations.

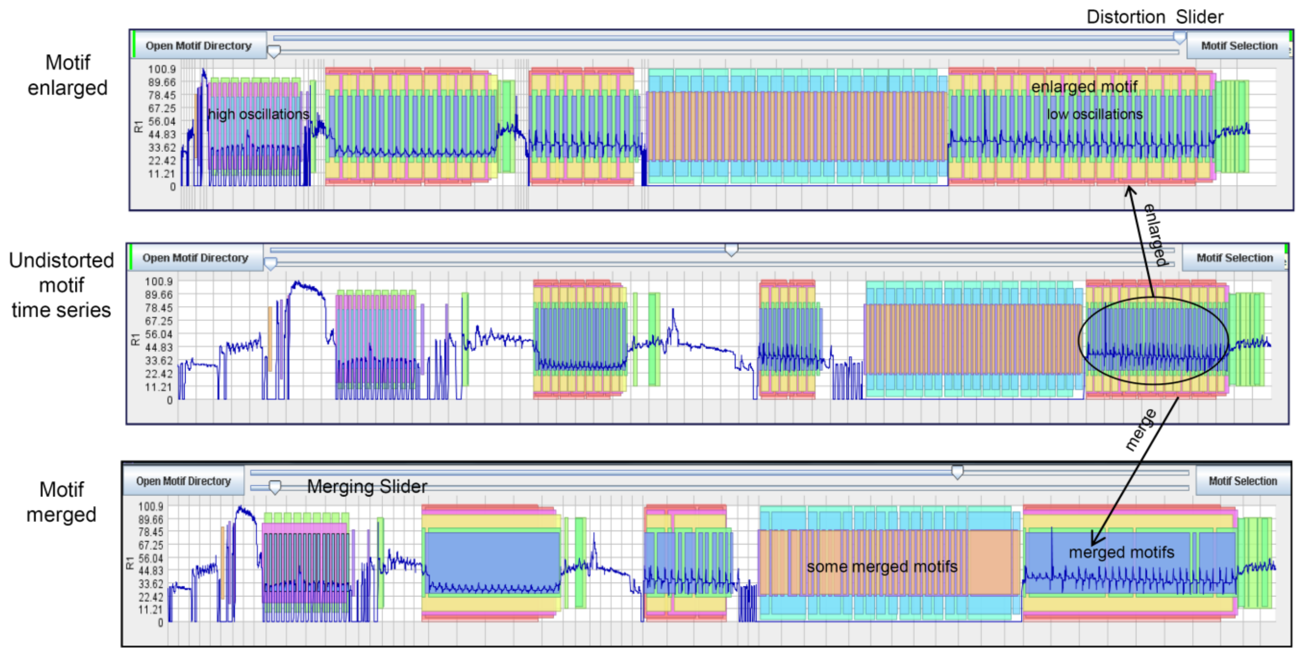


Figure 3: Motif Visual Distortion and Merging
(x-axis: time intervals, y-axis: %utilization of chiller R1, rectangles: motifs, color: motif types)
If the distortion slider is moved to the right, motifs are enlarged. If the merging slider is moved to the right, adjacent motifs of the same type are merged.

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:

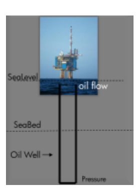
Distortion*:* Distortion enlarges the selected motifs using a user-activated distortion slider shown in the top of motif of Figure 3. Distorting the time series is done by applying an adapted density-equalizing distortion technique. We calculate weights for each time interval and use them as the input 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 the above techniques to analyze data center chiller sensor time series and oil well production sensor data (e.g., oil flow, pressure). Figure 3 provides an enlarged view of the motifs, allowing data center administrators to compare them in terms of their oscillatory behavior. They prefer not to operate chillers in motifs exhibiting high frequency oscillations since that decreases the life span of the chillers.



Oil Well Production

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] 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.

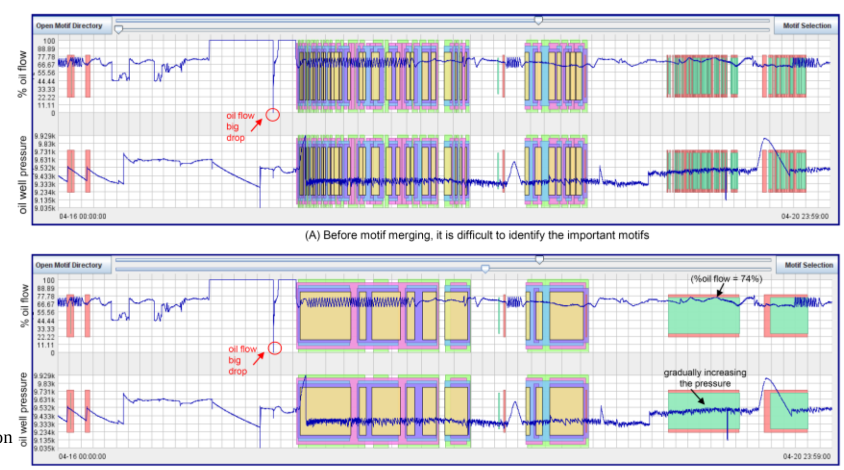


Figure 4: Oil Well Production Time Series (85,035 records) with Seven Different Motifs

(x-axis: time, y-axis: % oil flow and pressure, color: motif type)